Bayesian strong gravitational-lens modelling on adaptive grids

Vegetti, S.; Koopmans, L. V. E.

Published in:
Monthly Notices of the Royal Astronomical Society

DOI:
10.1111/j.1365-2966.2008.14005.x

Publication date:
2009

Mon. Not. R. Astron. Soc. (2008) doi:10.1111/j.1365-2966.2008.14005.x

Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in Galaxies

S. Vegetti⋆ and L. V. E. Koopmans
Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, the Netherlands

Accepted 2008 September 24. Received 2008 September 15; in original form 2008 May 5

ABSTRACT

We introduce a new adaptive and fully Bayesian grid-based method to model strong gravitational lenses with extended images. The primary goal of this method is to quantify the level of luminous and dark mass substructure in massive galaxies, through their effect on highly magnified arcs and Einstein rings. The method is adaptive on the source plane, where a Delaunay tessellation is defined according to the lens mapping of a regular grid on to the source plane. The Bayesian penalty function allows us to recover the best non-linear potential-model parameters and/or a grid-based potential correction and to objectively quantify the level of regularization for both the source and potential. In addition, we implement a Nested-Sampling technique to quantify the errors on all non-linear mass model parameters – marginalized over all source and regularization parameters – and allow an objective ranking of different potential models in terms of the marginalized evidence. In particular, we are interested in comparing very smooth lens mass models with ones that contain mass substructures. The algorithm has been tested on a range of simulated data sets, created from a model of a realistic lens system. One of the lens systems is characterized by a smooth potential with a power-law density profile, 12 include a Navarro, Frenk and White (NFW) dark matter substructure of different masses and at different positions and one contains two NFW dark substructures with the same mass but with different positions. Reconstruction of the source and lens potential for all of these systems shows the method is able, in a realistic scenario, to identify perturbations with masses $\gtrsim 10^7\,\mathrm{M}_\odot$ when located *on* the Einstein ring. For positions both inside and outside of the ring, masses of at least $10^9\,\mathrm{M}_\odot$ are required (i.e. roughly the Einstein ring of the perturber needs to overlap with that of the main lens). Our method provides a fully novel and objective test of mass substructure in massive galaxies.

Key words: gravitational lensing – galaxies: haloes – galaxies: structure – dark matter.

1 INTRODUCTION

At the present time, the most popular cosmological model for structure formation is the Λ cold dark matter (ΛCDM) paradigm. While this model has been very successful in describing the Universe on large scales and in reproducing numerous observational results (e.g. Reiss et al. 1998; Burles, Nollett & Turner 2001; Jaffe et al. 2001; Percival et al. 2001; Phillips et al. 2001; Croft et al. 2002; de Bernardis et al. 2002; Efstathiou et al. 2002; Hamilton & Tegmark 2002; Spergel et al. 2003; Tonry et al. 2003; Komatsu et al. 2008), important discrepancies still persist on small scales. In particular, some of these involve the dark matter distribution within galactic haloes (e.g. Moore 1994; Burkert 1995; McGaugh & de Blok 1998; Binney & Evans 2001; de Blok, McGaugh & Rubin 2001; de Blok & Bosma 2002; McGaugh, Barker & de Blok 2003; Simon et al. 2003; Rhee et al. 2004; Kuzio de Naray et al. 2006) and the number of galaxy satellites, i.e. the *Missing Satellite Problem*.

According to the standard scenario, structures form in a hierarchical fashion via merging and accretion of smaller objects (Toomre 1977; Frenk et al. 1988; White & Frenk 1991; Barnes 1992; Cole et al. 2000). As shown by the latest numerical simulations, in which high mass and force resolution is achieved, the progenitor population is only weakly affected by virialization processes and a large number of subhaloes are able to survive after merging. The number of substructures within the Local Group, however, is predicted to be one to two orders of magnitude higher than what is effectively observed (e.g. Kauffmann, White & Guiderdoni 1993; Klypin et al. 1999; Moore et al. 1999, 2001; Diemand, Kuhlen & Madau 2007a,b).

Two different classes of solutions have been suggested to alleviate this problem, cosmological and astrophysical. Cosmological

⋆E-mail: vegetti@astro.rug.nl

solutions address the basis of the ΛCDM paradigm itself and mostly concentrate on the properties of the dark matter, allowing for example, for a warm (Colin, Avila-Reese & Valenzuela 2000), decaying (Cen 2001), self-interacting (Spergel & Steinhardt 2000), repulsive (Goodman 2000) or annihilating nature (Riotto & Tkachev 2000). Alternatively, the ΛCDM picture can be modified by the introduction of a break of the power spectrum at the small scales (e.g. Kamionkowski & Liddle 2000; Zentner & Bullock 2003).

From an astrophysical point of view, the number of visible satellites can be reduced by suppressing the gas collapse/cooling (e.g. Bullock, Kravtsov & Weinberg 2000; Kravtsov, Gnedin & Klypin 2004; Moore et al. 2006) via supernova feedback, photoionization or reionization. This would result in a high mass-to-light ratio (M/L) in the substructures. If these high-M/L substructures indeed exist, different methods for indirect detection are possible. The dark substructure may be detectable, for example, through its effects on stellar streams (e.g. Ibata et al. 2002; Mayer et al. 2002), via γ-rays from dark matter annihilation (Bergström et al. 1999; Calcáneo-Roldán & Moore 2000; Stoehr et al. 2003; Colafrancesco, Profumo & Ullio 2006) or through gravitational lensing (e.g. Dalal & Kochanek 2002; Koopmans 2005).

While the first two approaches are limited to the local Universe, gravitational lensing allows one to explore the mass distribution of galaxies outside the Local Group and at a relatively high redshift. Moreover, gravitational lensing is independent of the baryonic content, the dynamical state of the system and the nature of dark matter. For example, when in a lens system a point source is close to the caustic fold or cusp, the sum of the image fluxes should add to zero if the sign of the image parities is taken into account (Blandford & Narayan 1986; Zakharov 1995). This relation is, however, violated by many observed lensed quasars with cusp and fold images. As first suggested by Mao & Schneider (1998), these flux ratio anomalies can be related to the presence of (dark matter) substructure around the lensing galaxy on scales smaller than the image separation (Bradač et al. 2002; Chiba 2002; Dalal & Kochanek 2002; Metcalf & Zhao 2002; Keeton, Gaudi & Petters 2003; Bradač et al. 2004; Kochanek & Dalal 2004; Keeton, Gaudi & Petters 2005). Nevertheless, subsequent studies of similar gravitationally lensed systems have shown that the required mass fraction in substructure is higher than what is obtained in numerical simulations (Mao et al. 2004; Macciò & Miranda 2006; Diemand et al. 2007a). In addition, for a significant number of cases, the observed flux ratio anomalies can be explained by taking into account the luminous dwarf satellite population (Ros et al. 2000; Trotter, Winn & Hewitt 2000; Koopmans & Treu 2002; Kochanek & Dalal 2004; Chen et al. 2007; McKean et al. 2007; More et al. 2008). Whether the mass fraction of CDM substructures is quantifiable via flux ratio anomalies is therefore a question still open for debate. Alternatively, Koopmans (2005) showed that dark matter substructure in lensing galaxies can be detected by modelling of multiple images or Einstein rings from extended sources.

In this paper, we developed an adaptive grid-based modelling code for extended lensed sources and grid-based potentials, to fully quantify this procedure. The method presented here is a significant improvement of the techniques introduced by Warren & Dye (2003), Dye & Warren (2005), Koopmans (2005), Suyu & Blandford (2006), Suyu et al. (2006) and Brewer & Lewis (2006). In order to detect mass substructure in lens galaxies, one needs to solve simultaneously for the source surface brightness distribution and the lens potential. A semilinear technique for the reconstruction of grid-based sources, given a parametric lens potential, was first introduced by Warren & Dye (2003). The method was subsequently extended by Koopmans (2005) and Suyu & Blandford (2006) in order to include a grid-based potential for the lens and by Barnabè & Koopmans (2007) to include galaxy dynamics. Dye & Warren (2005) introduced an adaptive gridding on the source plane; this would minimize the covariance between pixels and decrease the computational effort. However, the method is still lacking an objective procedure to quantify the level of regularization. Suyu et al. (2006) and Brewer & Lewis (2006) encoded the semilinear method within the framework of Bayesian statistics (MacKay 1992, 2003). Although a vast improvement, the fixed grids do not allow us to take into account the correct number of degrees of freedom and proper evidence comparison is difficult. In the implementation here described, these issues have been solved:

(i) the procedure is fully Bayesian; this allows us to determine the best set of non-linear parameters for a given potential and the linear parameters of the source, to objectively set the level of regularization and to compare/rank different potential families;

(ii) using a Delaunay tessellation, the source grid automatically adaptives in such a way that the computational effort is mostly concentrated in high-magnification regions;

(iii) the source-grid triangles are recomputed at every step of the modelling so that the source and the image plane always perfectly map on to each other and the number of degrees of freedom remains constant during Bayesian evidence maximization.

For the first time in the framework of grid-based lensing modelling, we use the Nested-Sampling technique by Skilling (2004) to compute the full marginalized Bayesian evidence of the data (MacKay 1992, 2003). This approach not only provides statistical errors on the lens parameters, but also consistently quantifies the relative evidence of a smooth potential against one containing substructures. As such, our method provides a fully objective way to rank these two hypotheses given the data, which is the goal set out in this paper.

The paper is organized as follow. In Section 2, we give a general overview on the data model. In Section 3, we present in detail how the data model can be inverted and the source and lens potential reconstructed. In Section 4, we review the basics of Bayesian statistics and Nested-Sampling technique for evidence computation. In Section 5, we describe how the method has been tested and how its ability in detecting substructures, depending on the perturbation mass and position, has been studied. Finally in Section 6, conclusions are drawn and future applications are discussed.

2 CONSTRUCTION OF THE LENSING OPERATORS

In this section, we describe the data model which relates the unknown source brightness distribution and lens potential to the known data of the lensed images. The aim is to put this procedure in a fully self-consistent mathematical framework, excluding as much as possible any subjective intervention into the modelling. The core of the method presented here is based on a Occam's razor argument. From a Bayesian evidence point of view, correlated features in the lensed images are most likely due to structure in the source, rather than being the result of small-scale perturbations of the lens potential in front of all the lensed images. On the other hand, uncorrelated structure in the lensed images is most likely due to small-scale perturbations of the lens potential.

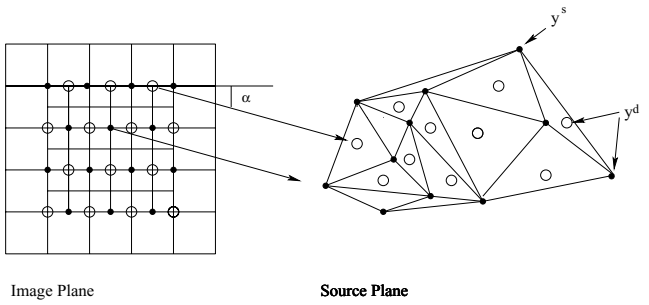

Figure 1. A schematic overview of the non-linear source and potential reconstruction method, as implemented in this paper. On the left-hand side, on the image plane, two grids are defined: one for the potential corrections and the other for the lensed image. A subset of N_s of the N_d image pixels located at the positions $\boldsymbol{x}_i^s$ on the image plane (filled circles) is cast back to the source plane (on the right-hand panel) on $\boldsymbol{y}_i^s$ through the lens equation. These form the vertices of an adaptive grid on the source plane. The remaining image pixels (open circles) are also cast to the source plane to the positions $\boldsymbol{y}_i^d$ (we note that this set of points includes $\boldsymbol{y}_i^s$). Because the source brightness distribution is conserved, i.e. $S(\boldsymbol{x}_i^d) = S(\boldsymbol{y}_i^d)$, the surface brightness at the empty circles is represented by a linear superposition of the surface brightness at the three triangle vertices that enclose it. Similarly, the potential correction at a point $\boldsymbol{x}_i^{\delta\psi}$ is given by a linear interpolation of the potential corrections at the surrounding pixels (large rectangular pixels on the image plane).

2.1 The data, source and potential grids

The main idea of grid-based lensing techniques is to use a grid-based reconstruction of the source and the lens potential. Here, we introduce the general geometry of the problem, explicitly shown in Fig. 1. Consider a lensed image $\boldsymbol{d}$ of an unknown extended source $\boldsymbol{s}$. Both $\boldsymbol{d}$ and $\boldsymbol{s}$ are vectors that describe the surface brightness distributions on a set of spatial points $\boldsymbol{x}_i^d$ and $\boldsymbol{y}_j^s$ in the lens and source plane, respectively (e.g. Warren & Dye 2003; Koopmans 2005; Suyu et al. 2006). In general, these are related through the lens equation $\boldsymbol{y}_i^d = \boldsymbol{x}_i^d - \nabla\psi(\boldsymbol{x}_i^d)$, where $\boldsymbol{x}_i^d$ corresponds to the spatial position of the surface brightness in the ith element of the vector $\boldsymbol{d}$, i.e. d_i and $\psi(\boldsymbol{x}_i^d)$ is the lensing potential, which is described in more detail in a moment. We note that $\boldsymbol{y}_i^d$ does not necessarily directly correspond to the elements $\boldsymbol{y}_j^s$, jth brightness value of the vector $\boldsymbol{s}$. In our implementation, the grid on the source plane is fully adaptive and directly constructed from a subset of the N_d pixels in the image plane, with spatial boundaries of the image grid included. In particular, as shown schematically in Fig. 1, N_s pixels, located each at a position $\boldsymbol{x}_i^s$ on the image grid, are cast back to the source plane giving the positions $\boldsymbol{y}_j^s$. The set of positions $\{\boldsymbol{y}_i^s\}$ constitutes the vertices of a Delaunay triangulation. In this way, we define an irregular adaptive grid, where vertex positions in the source plane are related to positions on the image plane via the lens equation and every vertex value represents an unknown source surface brightness level.

We assume the lens potential to be the superposition of a parametric smooth component with linear local perturbations related to the presence of, for example, CDM substructures or dwarf galaxies:

$$\psi(\boldsymbol{x}, \boldsymbol{\eta}) = \psi_s(\boldsymbol{x}, \boldsymbol{\eta}) + \delta\psi(\boldsymbol{x}). \tag{1}$$

While $\psi_s(\boldsymbol{x}, \boldsymbol{\eta})$ assumes a parametric form, with parameters $\boldsymbol{\eta}$, $\delta\psi(\boldsymbol{x})$ is a function that is pixelized on a regular Cartesian grid of points $\boldsymbol{x}_k^{\delta\psi}$ with values $\delta\psi_k$. The set $\{\delta\psi_k\}$ is written as a vector $\delta\boldsymbol{\psi}$. Given the observational set of data $\boldsymbol{d}$, we now wish to recover the source distribution $\boldsymbol{s}$ and the lens potential $\psi(\boldsymbol{x}, \boldsymbol{\eta})$ simultaneously. To do this, we need to mathematically relate the brightness values $\boldsymbol{d}$ to the unknown brightness values $\boldsymbol{s}$. As described in the next section, this can be done through a linear operation on $\boldsymbol{s}$ and $\delta\psi$, where the operator itself is a function of an initial guess of the lens potential.

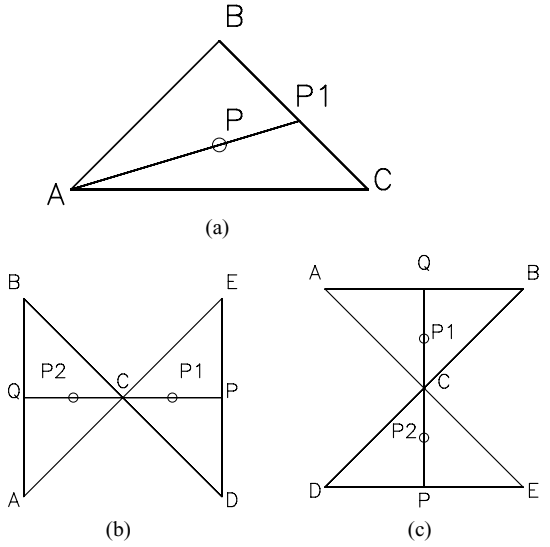

Figure 2. Generic triangles from the source grid. Both the source surface brightness and its derivatives at the points P, P_1 and P_2 are given by linear superposition of the values at the edges of the surrounding triangles.

2.2 The source and potential operator

We now derive the explicit relation between the unknown source distribution $\boldsymbol{s}$, the potential correction $\delta\boldsymbol{\psi}$, the smooth potential $\psi_s(\boldsymbol{x}, \boldsymbol{\eta})$ and the image brightness $\boldsymbol{d}$.

Consider a generic triangle $\widehat{\mathrm{ABC}}$ on the source plane (Fig. 2a), then the source surface brightness s_P on a point P, located inside the triangle at the position $\boldsymbol{y}_P^d$, can be related to the surface brightness on the vertices A, B and C through a simple linear relation:

$$s_P = w_A s_A + w_B s_B + w_C s_C. \tag{2}$$

An explicit expression for the bilinear interpolation weights w_A, w_B and w_C can be obtained by considering the point P_1, at the intersection of the line $\overline{\mathrm{AP}}$ with the line $\overline{\mathrm{CB}}$. The source intensities at P and P_1 are also related to each other through a linear interpolation. On the other hand, the surface brightness in P_1 is directly related to the values on the triangle vertices B and C:

$$\begin{cases} s_P = \frac{d_{PA}}{d_{P_1A}}(s_{P_1} - s_A) + s_A \\ s_{P_1} = \frac{d_{P_1B}}{d_{CB}}(s_C - s_B) + s_B \end{cases}, \tag{3}$$

where d_{PA} and d_{P_1A} are the absolute distances between the points P and A and the points P_1 and A, respectively; d_{P_1B} and d_{CB} are the distances between the points P_1 and B and the points C and B, respectively. Solving equation (3), we obtain the weights

$$\begin{cases} w_A = 1 - \frac{d_{PA}}{d_{P_1A}} \\ w_B = \frac{d_{PA}}{d_{P_1A}}\left(1 - \frac{d_{P_1B}}{d_{CB}}\right) \\ w_C = \frac{d_{PA}d_{P_1B}}{d_{P_1A}d_{CB}} \end{cases} \tag{4}$$

with $\sum_{i=A,B,C} w_i = 1$. Because gravitational lensing conserves the surface brightness, i.e. $S(\boldsymbol{x}_i^d) = S(\boldsymbol{y}_i^d)$, the mapping between the two

planes (when $\delta\boldsymbol{\psi} = 0$) can be expressed as a system of N_s coupled linear equations

$$\boldsymbol{B}\,\boldsymbol{L}(\boldsymbol{\eta})\boldsymbol{s} = \boldsymbol{d} + \boldsymbol{n}, \tag{5}$$

where $\boldsymbol{L}(\boldsymbol{\eta})$ and $\boldsymbol{B}$ are the lensing and the blurring operators, respectively (see e.g. Warren & Dye 2003; Treu & Koopmans 2004; Koopmans 2005; Suyu & Blandford 2006). The blurring operator is a square sparse matrix which accounts for the effects of the point spread function (PSF). Each row of the lensing operator (a sparse matrix) contains at most the three bilinear interpolation weights, $w_{\rm A,B,C}$, placed at the columns that correspond to the three source vertices that enclose the associated source position. For a vertex point, there is only one weight equal to unity. In case $N_s = N_d$ (i.e. all image positions are used to create the source grid), all weights are equal to unity. In this case, the systems of equations are underconstrained and strong regularization is required.

By pixelating $\delta\psi(\boldsymbol{x})$ on a regular Cartesian grid, a similar argument as for the source can be applied to the potential correction; all potential values, $\{\delta\psi_k\}$, and their derivatives on the image plane can be related to this limited set of points through bilinear interpolation (see Koopmans 2005; Suyu et al. 2008). It is then possible to derive from equation (5), a new set of linear equations:

$$\boldsymbol{M_c}\,(\boldsymbol{\eta}, \boldsymbol{\psi})\;\boldsymbol{r} = \boldsymbol{d} + \boldsymbol{n}, \tag{6}$$

where

$$\boldsymbol{r} \equiv \begin{pmatrix} \boldsymbol{s} \\ \delta\boldsymbol{\psi} \end{pmatrix}. \tag{7}$$

More specifically, $\boldsymbol{\psi}$ is the sum of all the previous corrections $\delta\boldsymbol{\psi}$ and the operator $\boldsymbol{M_c}$ is a block matrix reading

$$\boldsymbol{M_c} \equiv \boldsymbol{B}\left[\boldsymbol{L}(\boldsymbol{\eta}, \boldsymbol{\psi})\,|\,-\boldsymbol{D_s}(\boldsymbol{s}_{\rm MP})\boldsymbol{D}_{\boldsymbol{\psi}}\right]. \tag{8}$$

$\boldsymbol{L}(\boldsymbol{\eta}, \boldsymbol{\psi})$ is the lensing operator introduced above, $\boldsymbol{D}_s(\boldsymbol{s}_{\rm MP})$ is a sparse matrix whose entries depend on the surface brightness gradient of the previously best source model at $\boldsymbol{y}_i^d$ and $\boldsymbol{D}_\psi$ is a matrix that determines the gradient of $\delta\boldsymbol{\psi}$ at all corresponding points $\boldsymbol{x}_i^d$ (see Koopmans 2005, for details). The generic structure of these matrices is given by

$$\boldsymbol{D_s} = \begin{pmatrix} \cdots & & & & \\ & \frac{\partial S(\boldsymbol{y}_i^d)}{\partial y_1} & \frac{\partial S(\boldsymbol{y}_i^d)}{\partial y_2} & & \\ & & \frac{\partial S(\boldsymbol{y}_{i+1}^d)}{\partial y_1} & \frac{\partial S(\boldsymbol{y}_{i+1}^d)}{\partial y_2} & \\ & & & & \cdots \end{pmatrix} \tag{9}$$

and

$$\boldsymbol{D}_{\delta\boldsymbol{\psi}} = \begin{pmatrix} \cdots & & & & \\ & \frac{\partial\delta\psi(\boldsymbol{x}_i^d)}{\partial x_1} & & & \\ & \frac{\partial\delta\psi(\boldsymbol{x}_i^d)}{\partial x_2} & & & \\ & & \frac{\partial\delta\psi(\boldsymbol{x}_{i+1}^d)}{\partial x_1} & & \\ & & \frac{\partial\delta\psi(\boldsymbol{x}_{i+1}^d)}{\partial x_2} & & \\ & & & & \cdots \end{pmatrix}, \tag{10}$$

where the index i runs along all the $\boldsymbol{x}_i^d$ and $\boldsymbol{y}_i^d$, i.e. triangle vertices included. The 'functions' S and $\delta\psi$ and their derivative can be derived through bilinear interpolation and finite differencing from $\boldsymbol{s}$ and $\delta\boldsymbol{\psi}$, respectively.

It is clear from the structure of these matrices that the first-order correction to the model, as a result of $\delta\psi$, is equal to $\delta d_i = -\nabla S(\boldsymbol{y}_i^d)\cdot\nabla\delta\psi(\boldsymbol{x}_i^d)$ at every point $\boldsymbol{x}_i^d$ (see e.g. Koopmans 2005, for a derivation).

As for the surface brightness itself, also the first derivatives for a generic point P on the source plane can be expressed as functions of the relative values on the triangle vertices A, B, C, yielding

$$\begin{aligned} \frac{\partial s_{\rm P}}{\partial y_1} &= w_{\rm A}\frac{\partial s_{\rm A}}{\partial y_1} + w_{\rm B}\frac{\partial s_{\rm B}}{\partial y_1} + w_{\rm C}\frac{\partial s_{\rm C}}{\partial y_1} \\ \frac{\partial s_{\rm P}}{\partial y_2} &= w_{\rm A}\frac{\partial s_{\rm A}}{\partial y_2} + w_{\rm B}\frac{\partial s_{\rm B}}{\partial y_2} + w_{\rm C}\frac{\partial s_{\rm C}}{\partial y_2}. \end{aligned} \tag{11}$$

For the generic vertex $j = $ A, B, C, these are given by $\frac{\partial s_j}{\partial y_1} = -\frac{n_0}{n_2}$ and $\frac{\partial s_j}{\partial y_2} = -\frac{n_1}{n_2}$, where $\boldsymbol{N} \equiv (n_0, n_1, n_2)$ is the unit-length surface normal vector at the vertex j and defined as the average of the adjacent per face normal vectors. For $\delta\boldsymbol{\psi}$ and its gradients, on a rectangular grid with rectangular pixels, we follow Koopmans (2005).

3 INVERTING THE DATA MODEL

As shown above, in both the cases of solving for the source alone, or solving for the source plus a potential correction, a *linear data model* can be constructed. In this section, we give a general overview of how this set of linear equations can be (iteratively) solved. A more thorough Bayesian description and motivation can be found in Section 4.

3.1 The penalty function

Before we go into the details of the method, we first restate that for a given lens potential $\psi(\boldsymbol{x}, \boldsymbol{\eta})$ and potential correction $\boldsymbol{\psi}_n = \sum_{i=1}^{n}\delta\boldsymbol{\psi}_i$, on a grid, the source surface brightness vector $\boldsymbol{s}$ and the data vector $\boldsymbol{d}$ can be related through a linear (matrix) operator:

$$\boldsymbol{M_c}(\boldsymbol{\eta}, \boldsymbol{\psi}_{n-1}, \boldsymbol{s}_{n-1})\boldsymbol{r}_n = \boldsymbol{d} + \boldsymbol{n}, \tag{12}$$

now explicitly written with their dependencies on the source and potential and with

$$\boldsymbol{r}_n = \begin{pmatrix} \boldsymbol{s}_n \\ \delta\boldsymbol{\psi}_n \end{pmatrix}. \tag{13}$$

In this equation, $\boldsymbol{s}_n$ is a model of the source brightness distribution at a given iteration n (we describe the iterative scheme momentarily). We assume the noise $\boldsymbol{n}$ to be Gaussian which is a good approximation for the *Hubble Space Telescope* (*HST*) images the method will be applied to. Even in case of deviations from Gaussianity, the central limit theorem, for many data points, ensures that the probability density distribution is often well approximated by a Normal distribution.

Because of the ill-posed nature of this relation, equation (12) cannot simply be inverted. Instead a penalty function which expresses the mismatch between the data and the model has to be defined by

$$P(\boldsymbol{s}, \delta\boldsymbol{\psi}\,|\,\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{s}_{n-1}, \boldsymbol{\psi}_{n-1}) = \chi^2 + \lambda_s^2\|\boldsymbol{H}_s\boldsymbol{s}\|_2^2 + \lambda_{\delta\psi}^2\|\boldsymbol{H}_{\delta\boldsymbol{\psi}}\delta\boldsymbol{\psi}\|_2^2 \tag{14}$$

with

$$\chi^2 = [\boldsymbol{M_c}(\boldsymbol{\eta}, \boldsymbol{\psi}_{n-1}, \boldsymbol{s}_{n-1})\,\boldsymbol{r} - \boldsymbol{d}]^T\,\boldsymbol{C_d^{-1}}\,[\boldsymbol{M_c}(\boldsymbol{\eta}, \boldsymbol{\psi}_{n-1}, \boldsymbol{s}_{n-1})\,\boldsymbol{r} - \boldsymbol{d}]. \tag{15}$$

The second and third terms in the penalty function contain prior information, or beliefs about the smoothness of the source and the potential, respectively, and $\boldsymbol{C_d}$ is the diagonal covariance matrix of the data. The level of regularization is set by the regularization parameters $\boldsymbol{\lambda}$, one for the source and other for the potential (see Koopmans 2005; Suyu et al. 2006, for a more general discussion). In a Bayesian framework, this penalty function is related to the

posterior probability of the model given the data (see Section 4). In the following two sections, we describe how to solve for the linear and non-linear parameters of the penalty function (except for $\boldsymbol{\lambda}$, which is described in Section 4).

3.1.1 Solving for the linear parameters

The most probable solution, $\boldsymbol{r}_{\rm MP}$, minimizing the penalty function is obtained by solving the set of linear equations:

$$(\boldsymbol{M}_c^T \boldsymbol{C}_{\rm d}^{-1} \boldsymbol{M}_{\rm c} + \boldsymbol{R}^T \boldsymbol{R})\, \boldsymbol{r} = \boldsymbol{M}_c^T \boldsymbol{C}_{\rm d}^{-1} \boldsymbol{d}. \tag{16}$$

The regularization matrix is given by

$$\boldsymbol{R}^T \boldsymbol{R} = \begin{pmatrix} \lambda_{\rm s}^2 \boldsymbol{H}_s^T \boldsymbol{H}_s & \\ & \lambda_{\delta\psi}^2 \boldsymbol{H}_{\boldsymbol{\delta\psi}}^T \boldsymbol{H}_{\boldsymbol{\delta\psi}} \end{pmatrix}. \tag{17}$$

The solution of this symmetric positive definite set of equations can be found using, for example, a Cholesky decomposition technique. By solving equation (16), adding the correction $\delta\boldsymbol{\psi}_n$ to the previously best potential $\boldsymbol{\psi}_{n-1}$ and iterating this procedure, both the source and the potential should converge to the minimum of the penalty function $P(\boldsymbol{s}_n, \delta\boldsymbol{\psi}_n \,|\, \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{s}_{n-1}, \boldsymbol{\psi}_{n-1})$. At every step of this iterative procedure, the matrices $\boldsymbol{M_c}$ and $\boldsymbol{R}$ have to be recalculated for the new updated potential $\boldsymbol{\psi}_n$ and source $\boldsymbol{s}_n$. While the potential grid points are kept spatially fixed in the image plane, the Delaunay tessellation grid of the source is rebuilt at every iteration to ensure that the number of degrees of freedom is kept constant during the entire optimization process.

Note that because the source and the potential corrections are independent, they require their own form ($\boldsymbol{H}$) and level (λ) of regularization. The most common forms of regularization are the zeroth order, the gradient and the curvature. As shown by Suyu et al. (2006), the best form depends on the nature of the source distribution and can be assessed via Bayesian evidence maximization. For the source, we chose the curvature regularization defined for the Delaunay tessellation of the source plane.

Specifically, one can combine the gradient and curvature matrices in the x- and y-directions: $\boldsymbol{H}_s^T \boldsymbol{H}_s = \boldsymbol{H}_{s,y_1}^T \boldsymbol{H}_{s,y_1} + \boldsymbol{H}_{s,y_2}^T \boldsymbol{H}_{s,y_2}$. Both $\boldsymbol{H}_{s,y_1}$ and $\boldsymbol{H}_{s,y_2}$ can be obtained by analogy by considering the pair of triangles in Figs 2(b) and (c), respectively.

For every generic point C on the source plane, we consider the pair of triangles $\widehat{\rm ABC}$ and $\widehat{\rm DCE}$ and define the curvature in C in the y_1 direction as

$$s''_{{\rm C},y_1} \equiv \frac{1}{d_{\rm CP}}(s_{\rm P} - s_{\rm c}) - \frac{1}{d_{\rm CQ}}(s_{\rm c} - s_{\rm Q}). \tag{18}$$

This is not the second derivative, but we find that this alternative curvature definition gives much better results than using the second derivative directly. The reason is that it gives equal weight to all triangles, independently of their relative sizes (for identical rectangular pixels, this problem does not arise since the above definition is equal to the second derivative up to a proportionality constant). A much smoother solution in that case is obtained.

P and Q are given by intersecting the line $\overline{\rm CP_1}$ with the line $\overline{\rm ED}$ and the line $\overline{\rm CP_2}$ with the line $\overline{\rm AB}$, respectively. Specifically, $\rm P_1$ and $\rm P_2$ are defined as very small displacements from the point C in the y_1 direction:

$$\begin{aligned} y_2^{\rm P_1} &= y_2^{\rm P_2} = y_2^{\rm C} \\ y_1^{\rm P_{1,2}} &= y_1^{\rm C} \pm \delta y_1. \end{aligned} \tag{19}$$

The source surface brightness in P and Q can be obtained by linear interpolation between the source values in D with the value in E and the value in A with the value in B, respectively,

$$\begin{aligned} s_{\rm P} &= \frac{d_{\rm PD}}{d_{\rm ED}}(s_{\rm E} - s_{\rm D}) + s_{\rm D}, \\ s_{\rm Q} &= \frac{d_{\rm QA}}{d_{\rm AB}}(s_{\rm B} - s_{\rm A}) + s_{\rm A}\,, \end{aligned} \tag{20}$$

Substituting equation (20) in equation (18) gives

$$\begin{aligned} s''_{{\rm C},y_1} = &-\left(\frac{1}{d_{\rm CP}} + \frac{1}{d_{\rm CQ}}\right) s_{\rm C} + \frac{d_{\rm PD}}{d_{\rm CP} d_{\rm DE}} s_{\rm E} \\ &+ \frac{d_{\rm QA}}{d_{\rm CQ} d_{\rm AB}} s_{\rm B} + \frac{d_{\rm PE}}{d_{\rm CP} d_{\rm DE}} s_{\rm D} + \frac{d_{\rm QB}}{d_{\rm CQ} d_{\rm AB}} s_{\rm A}. \end{aligned} \tag{21}$$

Each row of the regularization matrix $\boldsymbol{H}_{s,y_1}$, corresponding to every point C, contains the five interpolation weights, placed at the columns that correspond to the five vertices A, B, C, D and E. The curvature in the y_2 direction is derived in an analogous way using the pair of triangles in Fig. 2(c). We refer again to Koopmans (2005) for details on the potential regularization matrix $H_{\delta\psi}$.

3.1.2 Solving for the non-linear parameters

In order to recover the non-linear parameters $\boldsymbol{\eta}$, we need to minimize the penalty function $P(\boldsymbol{s}, \boldsymbol{\eta} \,|\, \boldsymbol{\lambda}, \psi)$. We allow for a correction, $\boldsymbol{\psi}$, to the parametric potential $\psi(\boldsymbol{\eta}, \boldsymbol{x})$ (not necessarily zero), but do not allow it to be changed while optimizing for $\boldsymbol{s}$ and $\boldsymbol{\eta}$. In all the cases, we keep $\boldsymbol{\lambda}$ fixed during the optimization. Given an initial guess for the non-linear parameters $\boldsymbol{\eta}_0$, we then minimize the penalty function defined in Section 3.1.1, under the conditions outlined above ($\boldsymbol{\psi}$ is constant and $\delta\boldsymbol{\psi} \equiv 0$). We use a non-linear optimizer (in our case Downhill–Simplex with Simulated Annealing; Press et al. 1992) to change $\boldsymbol{\eta}$ at every step and to minimize the joint penalty function $P(\boldsymbol{s}, \boldsymbol{\eta} \,|\, \boldsymbol{\lambda}, \boldsymbol{\psi})$. The optimization of $\boldsymbol{s}$ is implicitly embedded in the optimization of $\boldsymbol{\eta}$ by solving equation (16) only for $\boldsymbol{s}$, every time $\boldsymbol{\eta}$ is modified.

3.2 The optimization strategy

We have implemented a multifold optimization scheme for solving the linear equation (12). This scheme is not unique, but stabilizes the numerical optimization of this rather complex set of equations. Solving all parameters simultaneously would be computationally prohibitive and usually shows poor convergence properties.

3.2.1 Optimization steps

Our optimization scheme is similar to a *line-search* optimization, where consecutively different sets of unknown parameters are being kept fixed, while the others are optimized for. The sets $\{\delta\boldsymbol{\psi}, \boldsymbol{s}\}$, $\{\boldsymbol{\eta}, \boldsymbol{s}\}$ and $\{\boldsymbol{\lambda}, \boldsymbol{s}\}$ define the three different groups of parameters, of which only one is solved for at once. The individual steps, in no particular order, are then the following.

(i) We assume $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ to be constant vectors and iteratively solve for $\delta\boldsymbol{\psi}$ and the source $\boldsymbol{s}$. In this case, at every iteration we solve for $\boldsymbol{r}$ and adjust $\boldsymbol{\psi}$, using the linear correction to the potential $\delta\boldsymbol{\psi}$. This was described in Section 3.1.1.

(ii) We assume $\boldsymbol{\psi}$ and $\boldsymbol{\lambda}$ to be constant vectors and $\delta\boldsymbol{\psi}_i = 0$ at every iteration and only solve for the non-linear potential parameters $\boldsymbol{\eta}$ and the source $\boldsymbol{s}$. This was described in Section 3.1.2. We note that part of step (i) is also implicitly carried out in step (ii) (i.e. solving for $\boldsymbol{s}$).

(iii) We assume both (i) and (ii), above, and solve for the regularization parameters λ_s of the source and the source itself $\boldsymbol{s}$. This requires a Bayesian approach and will be described in more detail in Section 4. We have not attempted to optimize for $\lambda_{\delta\psi}$, but will study this in future publications.

The overall goal, however, remains to solve for the *full* set of unknown parameters $\{\boldsymbol{\eta}, \boldsymbol{\psi}_n, \boldsymbol{s}_n\}$ for $n \rightarrow \infty$ (or some large number). In particular, if an overall smooth (on scales of the image separations) potential model $\psi(\boldsymbol{\eta})$ does not allow a proper reconstruction of the lens system, we add an additional and a more flexible potential correction $\delta\boldsymbol{\psi}$, which can describe a more complex mass structure.

3.2.2 *Line-search optimization scheme*

In practice, we find that the optimal strategy to minimize the penalty function is the following, in order.

(i) We set λ_s to a large constant value such that the source model remains relatively smooth throughout the optimization (i.e. the peak brightness of the model is a factor of a few below that of the data) and keep $\boldsymbol{\psi}_n = 0$ (see also Suyu et al. 2006). We then solve for $\boldsymbol{\eta}$ and $\boldsymbol{s}$ that minimize the penalty function.

(ii) Once the best $\boldsymbol{\eta}$ and $\boldsymbol{s}$ are found, a Bayesian approach is used to find the best value of λ_s for the source only. At this point, $\boldsymbol{\psi}$ is still kept equal to zero.

(iii) Given the new value of λ_s, step (i) is repeated to find improved values of $\boldsymbol{\eta}$ and $\boldsymbol{s}$. Since the sensitivity of λ_s to changes in $\boldsymbol{\eta}$ is rather weak, at this point the best values of $\boldsymbol{\eta}$, $\boldsymbol{s}$ and $\boldsymbol{\lambda}$ have been found.

(iv) Next, all the above parameters are kept fixed and we solve for $\boldsymbol{r}$, this time assuming a very large value for $\lambda_{\delta\psi}$ to keep the potential correction (and convergence) smooth. We adjust $\boldsymbol{\psi}$ at every iteration until convergence is reached (e.g. Suyu et al., in preparation). At this point, we stop the optimization procedure.

(v) The smooth model with $\boldsymbol{\psi} = 0$ and the same model with $\boldsymbol{\psi} \neq 0$ is then compared through their Bayesian evidence values and errors on the parameters are estimated through the Nested Sampling of Skilling (2004) (Section 4).

Fig. 3 shows a complete flow diagram of our optimization scheme. In the next section, we place equation (14) and model ranking on a formal Bayesian footing. Those readers mostly interested in the application and tests of the method could continue reading in Section 5.

4 A BAYESIAN APPROACH TO DATA FITTING AND MODEL SELECTION

When trying to constrain the physical properties of the lens galaxy, within the grid-based approach, three different problems are faced. Given the linear relation in equation (6), we need to determine the linear parameters $\boldsymbol{r}$ for a certain set of data $\boldsymbol{d}$ and a form for the smooth potential $\psi_s(\boldsymbol{x}, \boldsymbol{\eta})$. We then aim to find the best values for the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ and finally, on a more general level, we wish to infer the best model for the overall potential and quantitatively rank different potential families. In particular, we want to compare smooth models with models that also include a potential grid for substructure (with more free parameters). These issues can all be quantitatively and objectively addressed within the framework of Bayesian statistics. In the context of data modelling, three levels of inference can be distinguished (MacKay 1992; Suyu et al. 2006).

(i) First level of inference: linear optimization. We assume the model $\boldsymbol{M_c}$, which depends on a given potential and source model, to be true and for a fixed form $\boldsymbol{R}$ and level ($\boldsymbol{\lambda}$) of regularization, we derive from Bayes' theorem the following expression:

$$P(\boldsymbol{r} \mid \boldsymbol{d}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{\eta}, \boldsymbol{M_c})\, P(\boldsymbol{r} \mid \boldsymbol{\lambda}, \boldsymbol{R})}{P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R})}. \quad (22)$$

The likelihood term, in case of Gaussian noise, for a covariance matrix $\boldsymbol{C_d}$ is given by

$$P(\boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{\eta}, \boldsymbol{M_c}) = \frac{1}{Z_d} \exp\left[-E_d(\boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{\eta}, \boldsymbol{M_c})\right], \quad (23)$$

where

$$Z_d = (2\pi)^{N_d/2} (\det \boldsymbol{C_d})^{1/2} \quad (24)$$

and (see equation 15)

$$E_d(\boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{\eta}, \boldsymbol{M_c}] = \frac{1}{2}\chi^2 = \frac{1}{2}(\boldsymbol{M_c r} - \boldsymbol{d})^T \boldsymbol{C}_d^{-1} (\boldsymbol{M_c r} - \boldsymbol{d}). \quad (25)$$

Because of the presence of noise and often the singularity of $\det(\boldsymbol{M_c}^T \boldsymbol{M_c})$, it is not possible to simply invert the linear relation in equation (6) but an additional penalty function must be defined through the introduction of a prior probability $P(\boldsymbol{r} \mid \boldsymbol{\lambda}, \boldsymbol{R})$ on $\boldsymbol{s}$

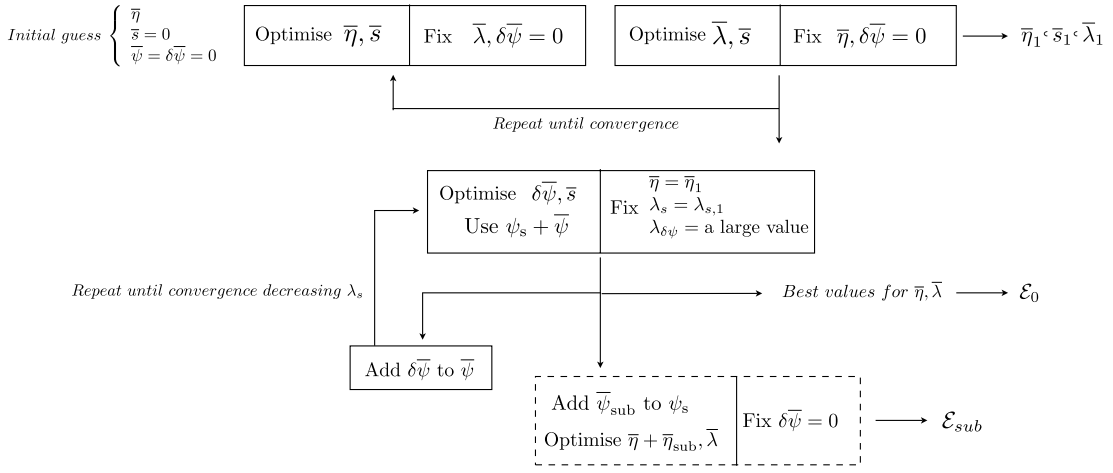

Figure 3. A schematic overview of the non-linear source and potential reconstruction method.

and $\delta\boldsymbol{\psi}$. In our implementation of the method, the prior assumes a quadratic form, with minimum in $\boldsymbol{r} = 0$ and sets the level of smoothness (specified in $\boldsymbol{H}$ and $\boldsymbol{\lambda}$) for the solution

$$P(\boldsymbol{r} \mid \boldsymbol{\lambda}, \boldsymbol{R}) = \frac{1}{Z_r} \exp\left[-\boldsymbol{\lambda} E_r(\boldsymbol{r} \mid \boldsymbol{R})\right], \tag{26}$$

with

$$Z_r(\boldsymbol{\lambda}) = \int \mathrm{d}\boldsymbol{r} \mathrm{e}^{-\boldsymbol{\lambda} E_r} = \mathrm{e}^{-\boldsymbol{\lambda} E_s(0)} \left(\frac{2\pi}{\boldsymbol{\lambda}}\right)^{N_r/2} (\det \boldsymbol{C})^{-1/2}, \tag{27}$$

$$E_r = \frac{1}{2} \|\boldsymbol{R}\boldsymbol{r}\|_2^2 \tag{28}$$

and

$$\boldsymbol{C} = \nabla\nabla E_r = \boldsymbol{R}\,\boldsymbol{R}^T. \tag{29}$$

The normalization constant $P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R})$ is called the evidence and plays an important role at higher levels of inference. In this specific case, it reads

$$P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) = \frac{\int \mathrm{d}\boldsymbol{r} \exp\left[-M(\boldsymbol{r})\right]}{Z_\mathrm{d} Z_r}, \tag{30}$$

where

$$M(\boldsymbol{r}) = E_\mathrm{d} + E_r. \tag{31}$$

The most probable solution for the linear parameters is found by maximizing the posterior probability:

$$P(\boldsymbol{r} \mid \boldsymbol{d}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) = \frac{\exp[-M(\boldsymbol{r})]}{\int \mathrm{d}\boldsymbol{r}\, \exp[-M(\boldsymbol{r})]}. \tag{32}$$

The condition $\partial(E_\mathrm{d} + E_r)/\partial\boldsymbol{r} = 0$ now yields the set of linear equations already introduced in Section 3.1.1:

$$\left(\boldsymbol{M_c}^\mathrm{T} \boldsymbol{C_d}^{-1} \boldsymbol{M_c} + \boldsymbol{R}^\mathrm{T}\boldsymbol{R}\right) \boldsymbol{r} = \boldsymbol{M_c}^\mathrm{T} \boldsymbol{C_d}^{-1} \boldsymbol{d}. \tag{33}$$

Equation (33) is solved iteratively using a Cholesky decomposition technique.

(ii) Second level of inference: non-linear optimization. At this level, we want to infer the non-linear parameters $\boldsymbol{\eta}$ and the hyperparameter λ_s for the source. Since at this point we are interested only in the smooth component of the lens potential, we set $\delta\boldsymbol{\psi} = 0$ and for a fixed family $\psi_\mathrm{s}(\boldsymbol{\eta})$, form of the regularization $\boldsymbol{R}$ and model $\boldsymbol{M_c}$, we maximize the posterior probability

$$P(\boldsymbol{\lambda}, \boldsymbol{\eta} \mid \boldsymbol{d}, \boldsymbol{M_c}, \boldsymbol{R}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) P(\boldsymbol{\lambda}, \boldsymbol{\eta})}{P(\boldsymbol{d} \mid \boldsymbol{M_c}, \boldsymbol{R})}. \tag{34}$$

Assuming a prior $P(\boldsymbol{\lambda}, \boldsymbol{\eta})$, which is flat in $\log(\lambda_\mathrm{s})$ and $\boldsymbol{\eta}$, reduces to maximizing the evidence $P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R})$ (which here plays the role of the likelihood) for $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$. The evidence can be computed by integrating over the posterior equation (34)

$$P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) = \int \mathrm{d}\boldsymbol{r}\, P(\boldsymbol{d} \mid \boldsymbol{r}, \boldsymbol{\eta}, \boldsymbol{M_c}) P(\boldsymbol{r} \mid \boldsymbol{\lambda}, \boldsymbol{R}). \tag{35}$$

Because of the assumptions we made (Gaussian noise and quadratic form of regularization), this integral can be solved analytically and yields

$$P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) = \frac{Z_M(\boldsymbol{\lambda}, \boldsymbol{\eta})}{Z_\mathrm{d} Z_r(\boldsymbol{\lambda})}, \tag{36}$$

where

$$Z_M(\boldsymbol{\lambda}, \boldsymbol{\eta}) = \exp\left[-M(\boldsymbol{r}_\mathrm{MP})\right] (2\pi)^{N_r/2} (\det \boldsymbol{A})^{-1/2}, \tag{37}$$

with $\boldsymbol{A} = \nabla\nabla M(\boldsymbol{r})$. Again we proceed in an iterative fashion: using a simulated annealing technique, we maximize the evidence (equation 35) for the parameters $\boldsymbol{\eta}$. Every step of the maximization generates a new model $\boldsymbol{M_c}[\psi(\boldsymbol{\eta}_i)]$, for which the most probable source $\boldsymbol{s}_\mathrm{MP}$ is reconstructed as described in Section (3). At this starting step, the level of the source regularization is set to a relatively large initial value $\lambda_{\mathrm{s},0}$; in this way, we ensure the solution to be smooth (at least at this first level) and the exploration of the $\boldsymbol{\eta}$ space to be faster. Subsequently, we fix the best model $\boldsymbol{M_c}(\boldsymbol{\eta}_0)$ found at the previous iteration and, using the same technique, we maximize the evidence for the source regularization level λ_s. The procedure is repeated until the total evidence has reached its maximum. In principle, we should have built a nested loop for λ_s at every step of the $\boldsymbol{\eta}$ exploration, but in practice the regularization constant only changes slightly with $\boldsymbol{\eta}$ and the alternate loop described above gives a faster way to reach the maximum (line-search method).

(iii) At the third level of inference, Bayesian statistics provides an objective and quantitative procedure for model comparison and ranking on the basis of the evidence:

$$P(\boldsymbol{M_c}, \boldsymbol{R} \mid \boldsymbol{d}) \propto P(\boldsymbol{d} \mid \boldsymbol{M_c}, \boldsymbol{R}) P(\boldsymbol{M_c}, \boldsymbol{R}). \tag{38}$$

For a flat prior $P(\boldsymbol{M_c}, \boldsymbol{R})$ (at this level of inference, we can make little to no assumptions) different models can be compared according to their value of $P(\boldsymbol{d} \mid \boldsymbol{M_c}, \boldsymbol{R})$, which is related to the evidence of the previous level by the following relation

$$P(\boldsymbol{d} \mid \boldsymbol{M_c}, \boldsymbol{R}) = \int \mathrm{d}\boldsymbol{\lambda}\, \mathrm{d}\boldsymbol{\eta}\, P(\boldsymbol{d} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{M_c}, \boldsymbol{R}) P(\boldsymbol{\lambda}, \boldsymbol{\eta}). \tag{39}$$

Being multidimensional and highly non-linear, the integral (equation 39) is carried out numerically through a Nested-Sampling technique (Skilling 2004), which is described in more detail in the next section. A byproduct of this method is an exploration of the posterior probability (equation 34), allowing for error analysis of the non-linear parameters and evidence itself.

4.1 Model selection: smooth versus clumpy models

In the previous section, we introduced the main structure of the Bayesian inference for model fitting and model selection. While parameter fitting simply determines how well a model matches the data and can be easily attained with the relatively simple analytic integrations of the first and second level of inference, model selection itself requires the highly non-linear and multidimensional integral (equation 39) to be solved. This marginalized evidence can be used to assign probabilities to models and reasonably establish whether the data require or allow additional parameters or not. Given two competing models M_0 and M_1 with relative marginalized evidence $\mathcal{E}_0$ and $\mathcal{E}_1$, the Bayes factor, $\Delta\mathcal{E} \equiv \log\mathcal{E}_0 - \log\mathcal{E}_1$, quantifies how well M_0 is supported by the data when compared with M_1 and it automatically includes the Occam's razor. Typically, the literature suggests to weigh the Bayes factor using Jeffreys' scale (Jeffreys 1961), which however provides only a qualitative indication: $\Delta\mathcal{E} < 1$ is not significant, $1 < \Delta\mathcal{E} < 2.5$ is significant, $2.5 < \Delta\mathcal{E} < 5$ is strong and $\Delta\mathcal{E} > 5$ is decisive.

In order to evaluate this marginalized evidence with a high enough accuracy, we implemented the new evidence algorithm known as Nested Sampling, proposed by Skilling (2004). Specifically, we would like to compare two different models: one in which the lens potential is smooth and other in which substructures are present, with, for example, a NFW profile. While the first is defined by the non-linear parameters of the lens potential and source regularization only, the second also allows us for three extra parameters: the mass of the substructure and its position on the lens plane (see Section 5)

4.2 Model ranking: nested sampling

Here, we provide a short description of how the Nested Sampling can be used to compute the marginalized evidence and errors on the model parameters; a more detailed one can be found in Skilling (2004). The Nested-Sampling algorithm integrates the likelihood over the prior volume by moving through thin nested likelihood surfaces. Introducing the fraction of total prior mass X, within which the likelihood exceeds $\mathcal{L}^*$, hence

$$X = \int_{\mathcal{L}>\mathcal{L}^*} \mathrm{d}X, \tag{40}$$

with

$$\mathrm{d}X = P(\boldsymbol{\lambda}, \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\lambda}\, \mathrm{d}\boldsymbol{\eta}, \tag{41}$$

the multidimensional integral (equation 39) relating the likelihood $\mathcal{L}$ and the marginalized evidence $\mathcal{E}$ can be reduced to a one-dimensional integral with positive and decreasing integrand

$$\mathcal{E} = \int_0^1 \mathrm{d}X\, \mathcal{L}(X), \tag{42}$$

where $\mathcal{L}(X)$ is the likelihood of the (possibly disjoint) iso-likelihood surface in parameter space which encloses a total prior mass of X. If the likelihood $\mathcal{L}_j = \mathcal{L}(X_j)$ can be evaluated for each of a given set of decreasing points, $0 < X_j < X_{j-1} < \ldots < 1$, then the total evidence $\mathcal{E}$ can be obtained, for example, with the trapezoid rule,

$$\mathcal{E} = \sum_{j=1}^{m} \mathcal{E}_j = \sum_{j=1}^{m} \frac{\mathcal{L}_j}{2}(X_{j-1} - X_{j+1}).$$

The power of the method is that the values of X_j do not have to be explicitly calculated, but can be statistically estimated. Specifically, the marginalized evidence is obtained through the following iterative scheme:

(i) the likelihood $\mathcal{L}$ is computed for N different points, called active points, which are randomly drawn from the prior volume;

(ii) the point X_j with the lowest likelihood is found and the corresponding prior volume is estimated statistically: after j iterations, the average volume decreases as $X_j/X_{j-1} = t$, where t is the expectation value of the largest of N numbers uniformly distributed between (0, 1);

(iii) the term $\mathcal{E}_j = \frac{\mathcal{L}_j}{2}\left(X_{j-1} - X_{j+1}\right)$ is added to the current value of the total evidence;

(iv) X_j is replaced by a new point randomly distributed within the remaining prior volume and satisfying the condition $\mathcal{L} > \mathcal{L}^* \equiv \mathcal{L}_j$;

(v) the above steps are repeated until a stopping criterion is satisfied.

By climbing up the iso-likelihood surfaces, the method, in general, find and quantifies the small region in which the bulk of the evidence is located.

Different stopping criteria can be chosen. Following Skilling (2004), we stop the iteration when $j \gg NH$, where H is minus the logarithm of that fraction of prior mass which contains the bulk of the posterior mass. In practical terms, this means that the procedure should be stopped only when most of the evidence has been included. Given the areas $\mathcal{E}_j$, in fact, the likelihood initially increases faster than the widths decrease, until its maximum is reached; across this maximum, located in the region $\mathcal{E} \approx \mathrm{e}^{-H}$, the likelihood flatten off and the decreasing widths dominate the increasing $\mathcal{L}_j$. Since $\mathcal{E}_j \approx \mathrm{e}^{-j/N}$, it takes NH iterations to reach the dominating areas. These NH iterations are random and subjected to a standard deviation uncertainty $\sqrt{NH}$, corresponding to a deviation standard on the logarithmic evidence of $\sqrt{NH}/N$:

$$\log \mathcal{E} = \log\left(\sum_j \mathcal{E}_j\right) \text{ with } \sigma_{\log \mathcal{E}} = \sqrt{\frac{H}{N}}. \tag{43}$$

4.2.1 Posterior probability distributions

For the lens parameters, the substructure position and the logarithm of the source regularization, priors are chosen to be uniform on a symmetric interval around the best values which we have determined at the second level of the Bayesian inference. The size of the interval being at least one order of magnitude larger than the errors on the parameters. In practice, we first carry out a fast run of the Nested Sampling with few active points N, this gives us an estimate for the non-linear parameter errors. Using the product $2 \times N_{\mathrm{dim}} \times \sigma_\eta$, where N_{dim} is the total number of parameters and σ_η is the corresponding standard deviation, we can then roughly enclose the bulk of the likelihood (note that this can be double-checked and corrected in hindsight, if the posterior probability functions are truncated at the prior boundaries). Priors on the parameters are taken in such a way that this maximum is fully included in the total integral of the marginalized evidence. For the main lens parameters and regularization constant, the same priors are used for model with and without substructure. For the substructure mass, a flat prior between $M_{\mathrm{min}} = 4.0 \times 10^6$ and $M_{\mathrm{max}} = 4.0 \times 10^9\, \mathrm{M}_\odot$ is adopted, with the two limits given by N-body simulations (e.g. Diemand et al. 2007a,b). In reality, the method does not require the parameters to be well known a priori, but limiting the exploration to the best-fitting region sensibly reduces the computational effort without significantly altering the evidence estimation. From Bayes theorem we have that the posterior probability density p_j is given by

$$p_j(t) = \frac{\mathcal{L}_j}{2}(X_{j-1} - X_{j+1})/\mathcal{E}(t) = w_j/\mathcal{E}(t). \tag{44}$$

The existing set of points $(\boldsymbol{\eta}, \boldsymbol{\lambda})_1, \ldots, (\boldsymbol{\eta}, \boldsymbol{\lambda})_N$ then gives us a set of posterior values that can be then used to obtain mean values and standard deviations on the non-linear parameters

$$\langle \boldsymbol{\eta} \rangle = \sum_j w_j \boldsymbol{\eta}_j / \sum_j w_j, \tag{45}$$

and similarly for $\boldsymbol{\lambda}$. These samples also provide a sampling of the full joint probability density function. Marginalizing over this function, the full marginalized probability density distribution of each parameters can be determined (see Section 5.5).

Table 1. Non-smooth (PL + NFW) lens models. At each of the P_i positions, a NFW perturbation of virial mass m_{sub} is superimposed on a smooth PL mass model distribution.

Lens	$\boldsymbol{x}_{\mathrm{sub}}$ (arcsec)	$m_{\mathrm{sub}}(\mathrm{M}_\odot)$
L_1	$P_0 = (+0.90; +1.19)$	10^7
L_2		10^8
L_3		10^9
L_4	$P_1 = (-0.50; -1.00)$	10^7
L_5		10^8
L_6		10^9
L_7	$P_2 = (-0.10; -0.60)$	10^7
L_8		10^8
L_9		10^9
L_{10}	$P_3 = (-0.90\,; -1.40)$	10^7
L_{11}		10^8
L_{12}		10^9
L_{13}	P_0 and P_1	10^8

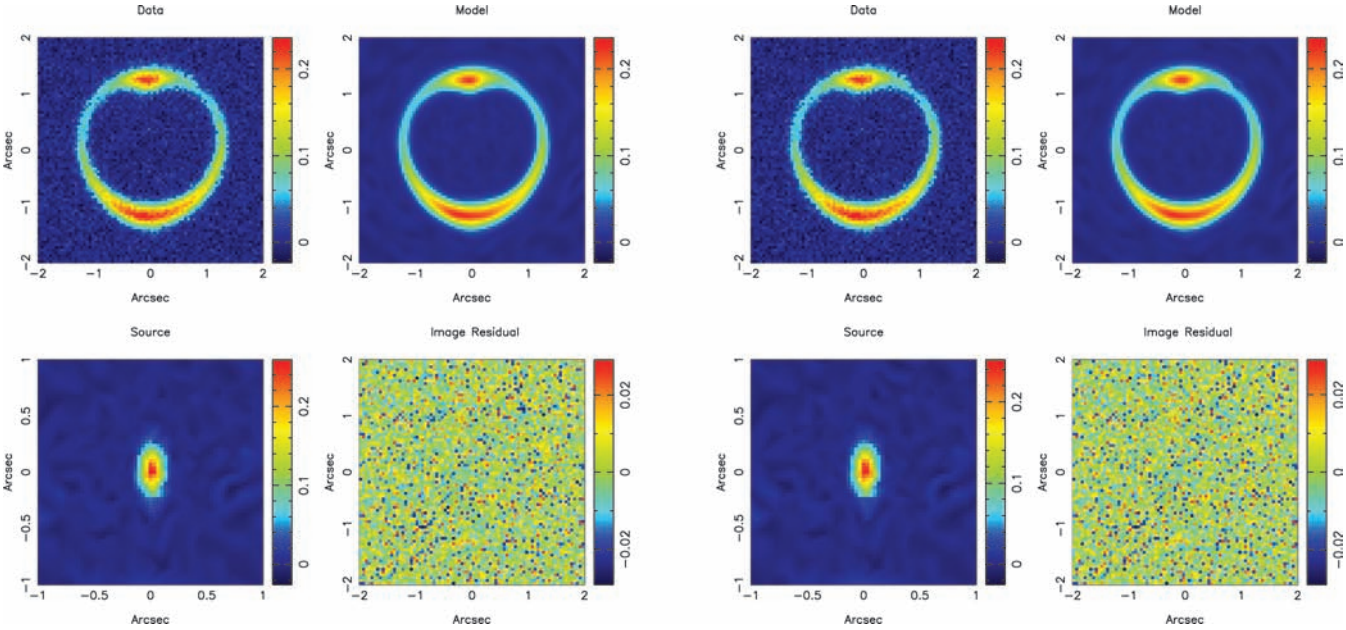

Figure 4. Left-hand panel: results of the first non-linear reconstruction for the smooth component of the perturbed lens L_1. The top right-hand panel shows the original mock data, while the top left-hand panel shows the final reconstruction. On the second row, the source reconstruction (left-hand panel) and the image residuals (right-hand panel) are shown. Right-hand panel: final results of the non-linear reconstruction for the perturbed lens L_1. The top right-hand panel shows the original mock data, while the top left-hand panel shows the final model reconstruction obtained after a non-linear optimization involving the lens parameters and the substructure position and mass. The recovered source is plotted in the low left-hand panel. Image residuals (right-hand panel) are shown.

5 TESTING AND CALIBRATING THE METHOD

In this section, we describe the procedure to test the method introduced above and to assess its ability to detect dark matter substructures in realistic data sets (e.g. from *HST*). A set of mock data, mimicking a typical Einstein ring, is created. We generate 14 different lens models, of which L_0 is purely smooth, $L_{1\leq i<13}$ are given by the superposition of the same smooth potential with a single NFW dark matter substructure of varying mass and position and L_{13} contains two NFW dark matter substructures with the same mass but with different positions (See Table 1). A first approximate reconstruction of the source and lens potential is performed by recovering the best non-linear lens parameters $\boldsymbol{\eta}$ and the level of source regularization λ_s. These values are then used for the linear grid-based optimization, which provides initial values of the substructure position and mass. Three extra runs of the non-linear optimization are then performed to recover the best set ($\boldsymbol{\eta}_b$, $\lambda_{s,b}$) for the main lens and the best mass and position of the substructure (solely modelled with a NFW density profile). Finally by means of the Nested-Sampling technique described in Section 4.1, we compute the marginalized evidence, equation (39), for every model twice, once under the hypothesis of a smooth lens and once allowing for the presence of one or two extra mass substructures. Comparison between these two models allows us to assess whether the presence of substructure in the model improves the evidence despite the larger number of free parameters.

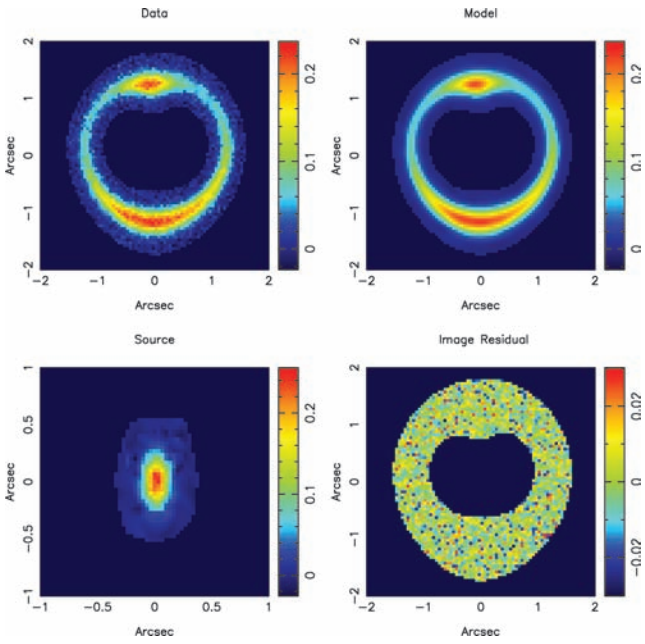

Figure 5. Results of the non-linear optimization for the smooth lens L_0. The top right-hand panel shows the original mock data, while the top left-hand panel shows the final reconstruction. On the second row, the source reconstruction (left-hand panel) and the image residuals (right-hand panel) are shown.

5.1 Mock data realizations

A set of simulated data with realistic noise is generated from a model based on the real lens Sloan Lens ACS Survey (SLACS) J1627−0055 (Bolton et al. 2006; Koopmans et al. 2006; Treu et al. 2006). We assume the lens to be well described by a power-law (PL) profile (Barkana 1998). Using the optimization technique described in Section 4, we find the best set of non-linear parameters ($\boldsymbol{\eta}_b$, $\lambda_{s,b}$). In particular, $\boldsymbol{\eta}$ contains the lens strength b, and some of the lens-geometry parameters: the position angle θ, the axis ratio f, the centre coordinates $\boldsymbol{x}_0$ and the density profile slope q, [$\rho \propto r^{-(2q+1)}$]. If necessary, information about external shear can be included. The best parameters are used to create 14 different lenses and their corresponding lensed images.

One of the systems is given by a smooth PL model while 12 include a dark matter substructure with virial mass $M_{vir} = 10^7, 10^8, 10^9\,M_\odot$ located either on the lowest surface brightness point of the ring P_0, on a high surface brightness point of the ring P_1, inside the ring P_2 and outside the ring P_3 (see Table 1). The 14 lens contains two substructures each with a mass of $M_{vir} = 10^8\,M_\odot$, located, respectively, in P_0 and P_1. The substructures are assumed to have a NFW profile:

$$\rho(r) = \rho_s(r_s/r)\,[1 + (r/r_s)]^{-2}, \tag{46}$$

where the concentration $c = r_{vir}/r_s$ and the scaling radius r_s are obtained from the virial mass using the empirical scaling laws provided by Diemand et al. (2007a,b). The source has an elliptical Gaussian surface brightness profile centred in zero

$$s\,(\boldsymbol{y}) = s_0 \exp\left[-(y_1/\delta y_1)^2 - (y_2/\delta y_2)^2\right]. \tag{47}$$

We assume $s_0 = 0.25$, $\delta y_1 = 0.01$ and $\delta y_2 = 0.04$.

5.2 Non-linear reconstruction of the main lens

We start by choosing an initial parameter set $\boldsymbol{\eta}_0$ for the main lens, which is offset from $\boldsymbol{\eta}_{true}$ that we used to create the simulated data. Assuming the lens does not contain any substructure, we run the non-linear procedure described in Section (4) and optimize $\{\boldsymbol{\eta}, \lambda_s\}$

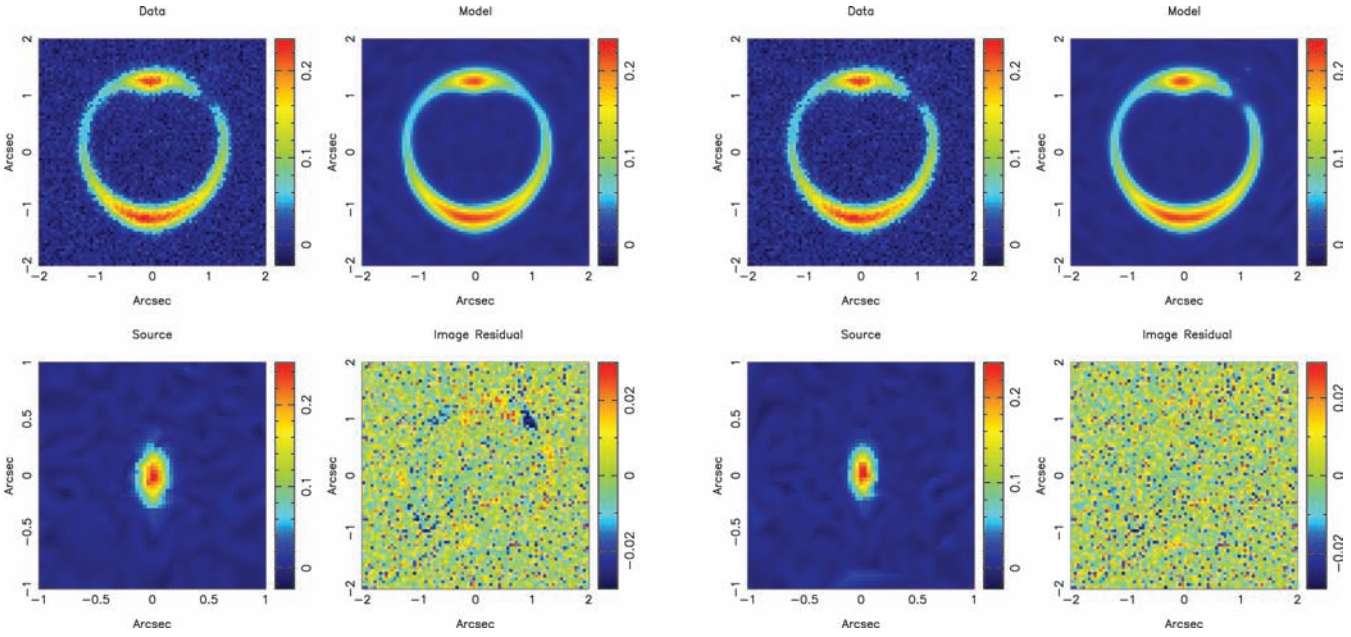

Figure 6. Similar as Fig. 4 for L_2.

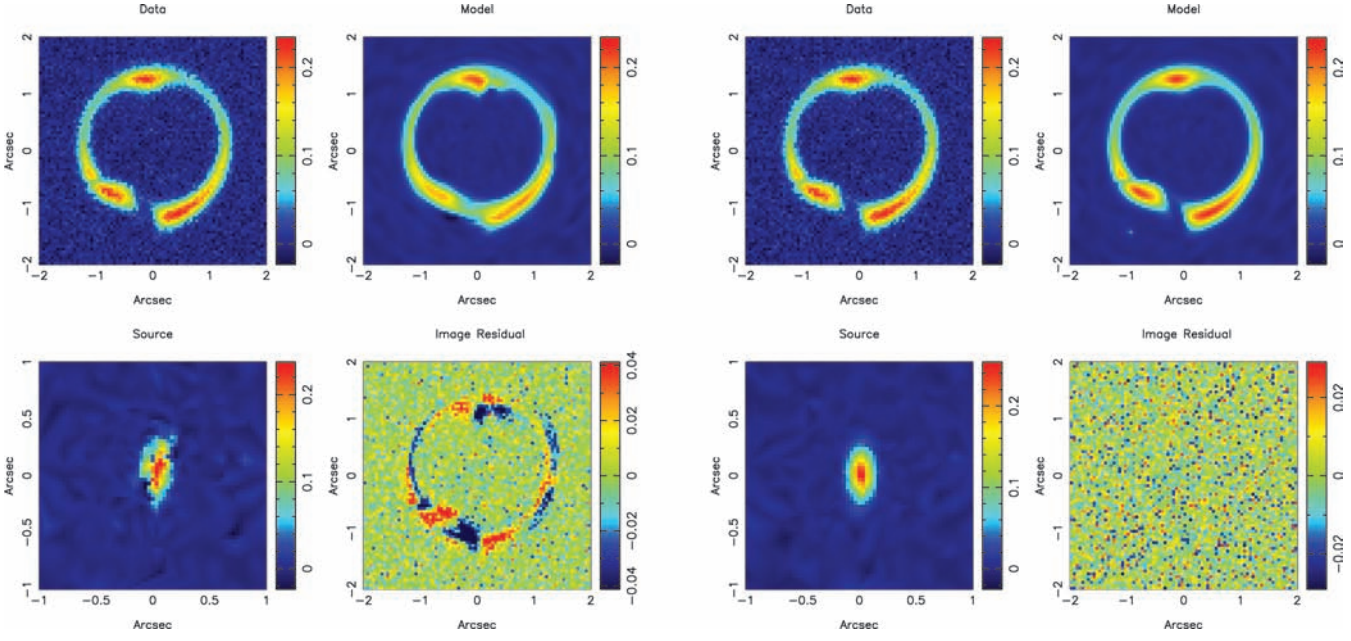

Figure 7. Similar as Fig. 4 for L_{12}.

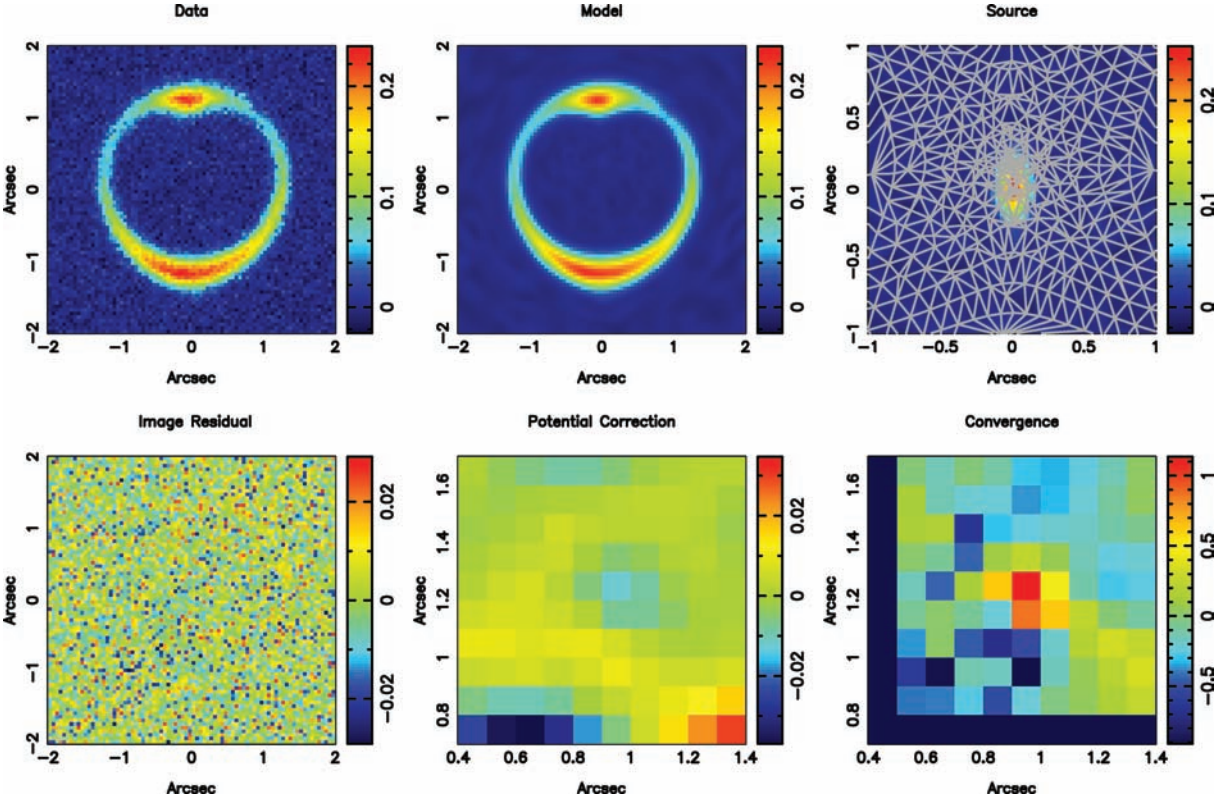

Figure 8. Results of the linear source and potential reconstruction for the lens L_1. The first row shows the original model (left-hand panel), the reconstructed model (middle panel) and the current-best source, as well as the corresponding adaptive grid. On the second row, the image residuals (left-hand panel), the total potential convergence (middle panel) and the substructure convergence (right-hand panel) are shown. Note that the substructure, although weak, is reconstructed at the correct position.

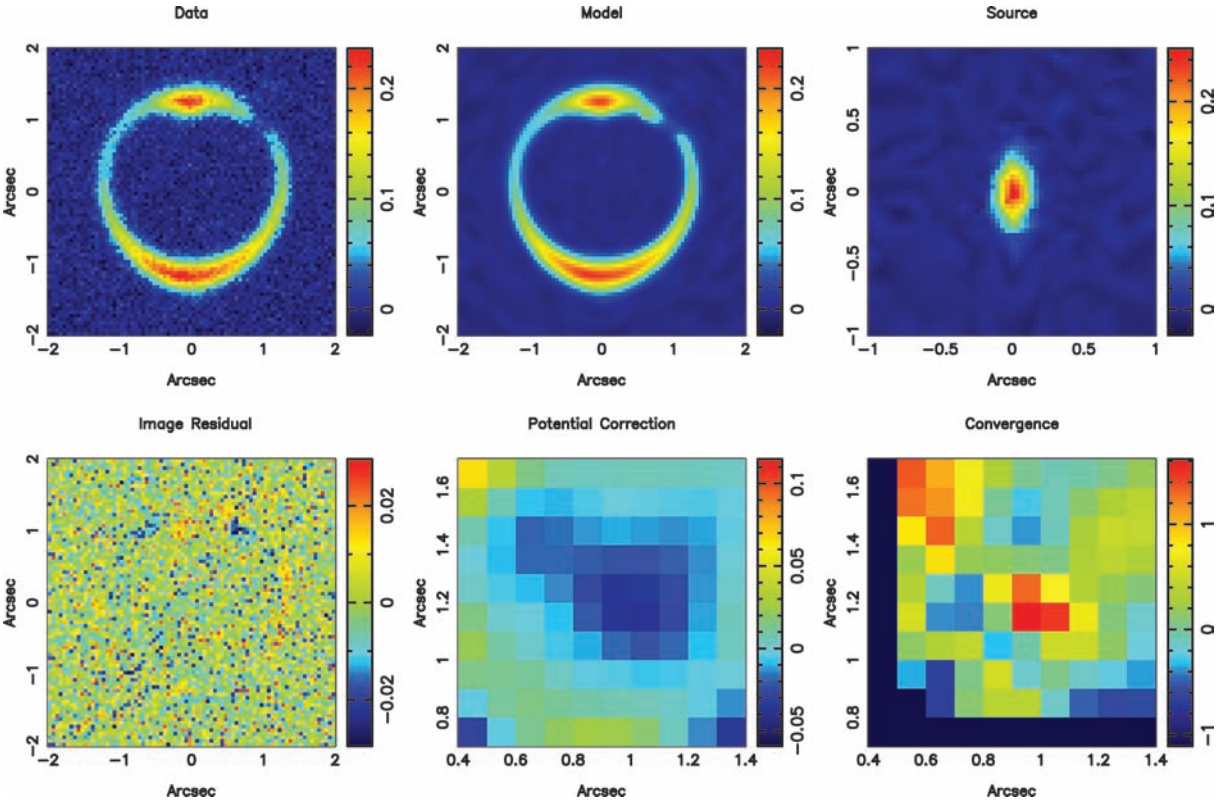

Figure 9. Similar as Fig. 8 for L_2. We note that the substructure is extremely well reconstructed, both at the correct position and in mass.

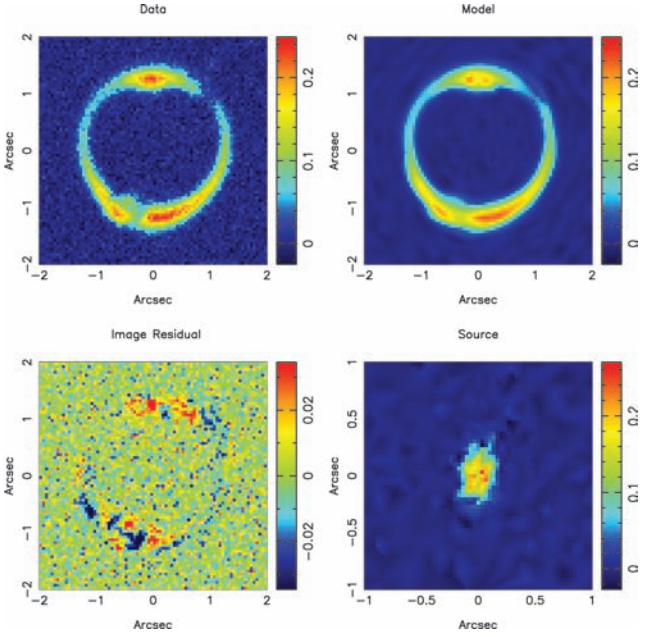

Figure 10. Non-linear reconstruction for the lens L_{13} for a single PL model.

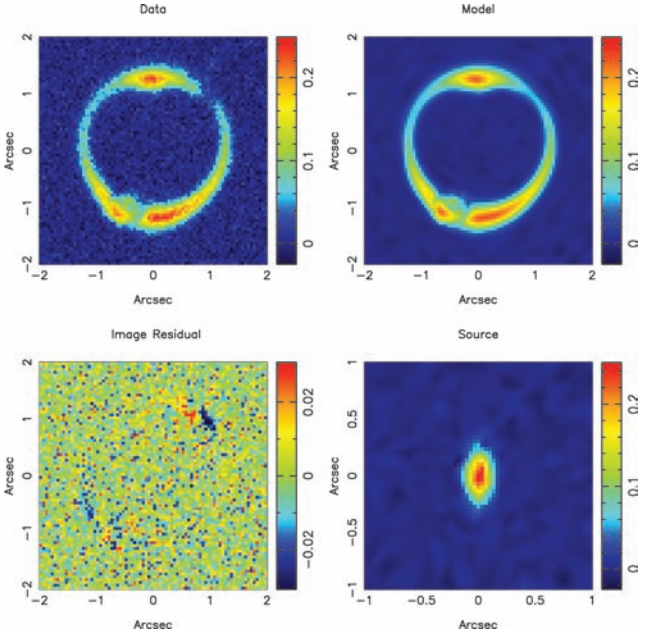

Figure 12. Non-linear reconstruction for the lens L_{13} for a PL + NFW model.

for each of the considered systems. At every step of the optimization, a new set $\{\boldsymbol{\eta}_i, \lambda_{s,i}\}$ is obtained and the corresponding lensing operator $\boldsymbol{M}_c(\boldsymbol{\eta}_i, \lambda_{s,i})$ has to be recomputed. The images are defined on a 81×81 pixels ($N_d = 6561$) regular Cartesian grid while the sources are reconstructed on a Delaunay tessellation grid of $N_s = 441$ vertices. The number of image points, used for the source-grid construction, is effectively a form of a prior and the marginalized evidence (equation 39) can be used to test this choice. To check whether the number of image pixels used can affect the result of our modelling, we consider the smooth lens L_0 and perform the non-linear reconstruction using 1 pixel every 16, 9, 4 and 1. In

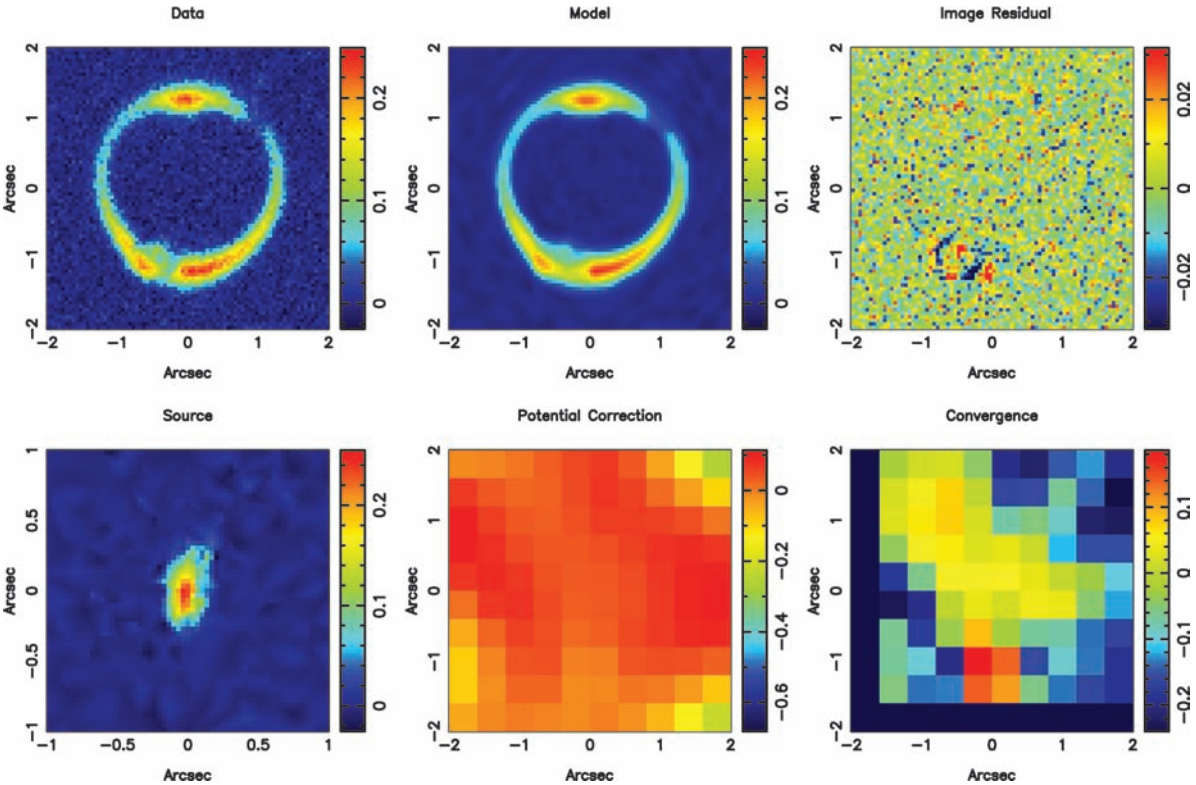

Figure 11. Results of the first linear source and potential reconstruction for the lens L_{13}. The first row shows the original model (left-hand panel), the reconstructed model (middle panel) and the image residuals (right-hand panel). On the second row, the current-best source (left-hand panel), the total potential convergence (middle panel) and the substructure convergence (right-hand panel) are shown. Note that the substructure, although weak, is reconstructed at the correct position.

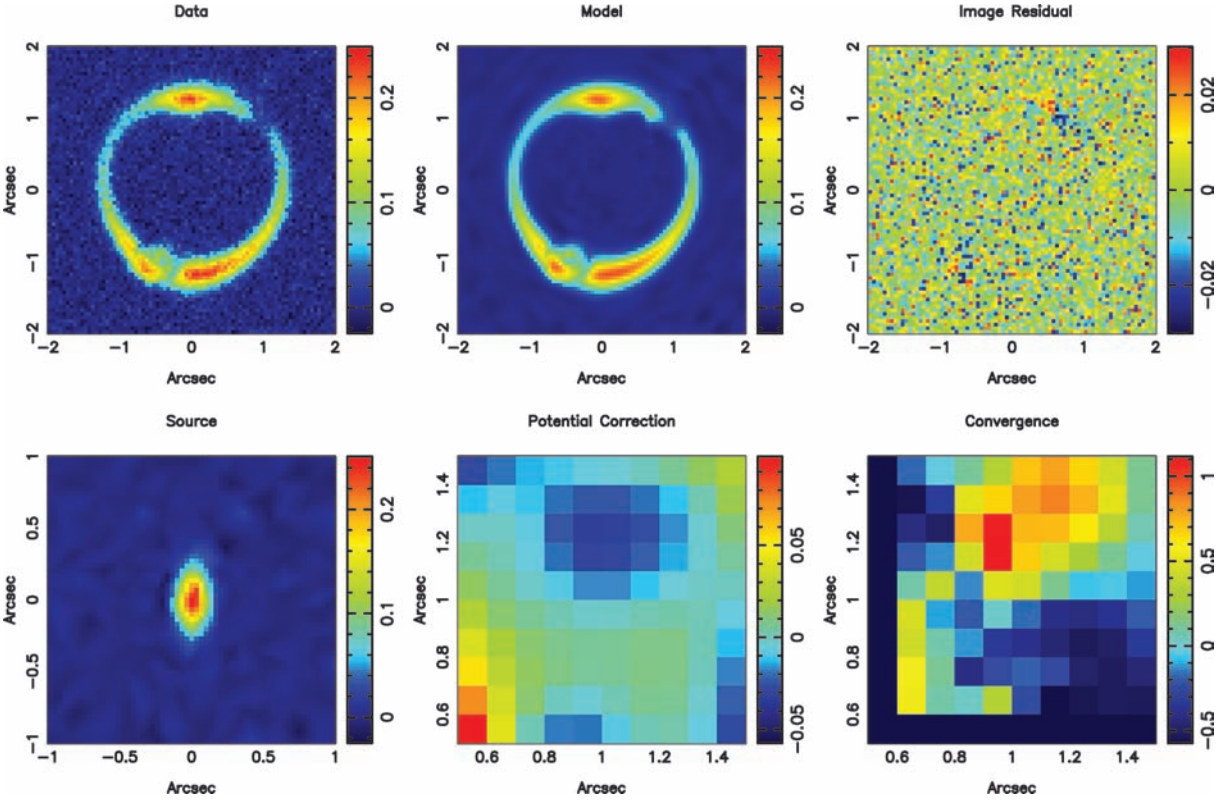

Figure 13. Results of the second linear source and potential reconstruction for the lens L_{13}.

each of the considered cases, we find that the lens parameters are within the relative errors (see Table 2). This suggests that, for this particular case, the choice of number of pixels is not influencing the quality of the reconstruction. In real systems, the dynamic range of the lensed images could be much higher and a case by case choice based on the marginalized evidence has to be considered. In Fig. 4, the residuals relative to the system L_1 are shown; the noise level is, in general, reached and only small residuals are observed at the position of the substructure. Whether the level of such residuals and therefore the relative detection of the substructure are significant is an issue we will address later on in terms of the total marginalized evidence.

5.3 Linear reconstruction: substructure detection

The non-linear optimization provides us with a first good approximate solution for the source and smooth component of the lens potential. While this is a good description for the smooth model L_0 (see Fig. 5), the residuals (e.g. Figs 6 and 7) for the perturbed model $L_{i\geq 1}$ indicate that the *no-substructure* hypothesis is improbable and perturbations to the main potential have to be considered. If the perturbation is small, this can be done by temporarily assuming that $\boldsymbol{\eta}_{i=1}$ reflects the true mass model distribution for the main lens and reconstruct the source and the potential correction by means of equation (33). In order to keep the potential corrections in the linear regime, where the approximation (equation 33) is valid, both the source and potential need to be initially over-regularized: $\lambda_s = 10\,\lambda_{s,1}$ and $\lambda_{\delta\psi} = 3.0 \times 10^5$ (Koopmans 2005; Suyu et al. 2006). For each of the possible substructure positions, we identify the lowest mass substructure we are able to recover. In the two most favourable cases, L_1 and L_4, in which the substructure sits on the Einstein ring

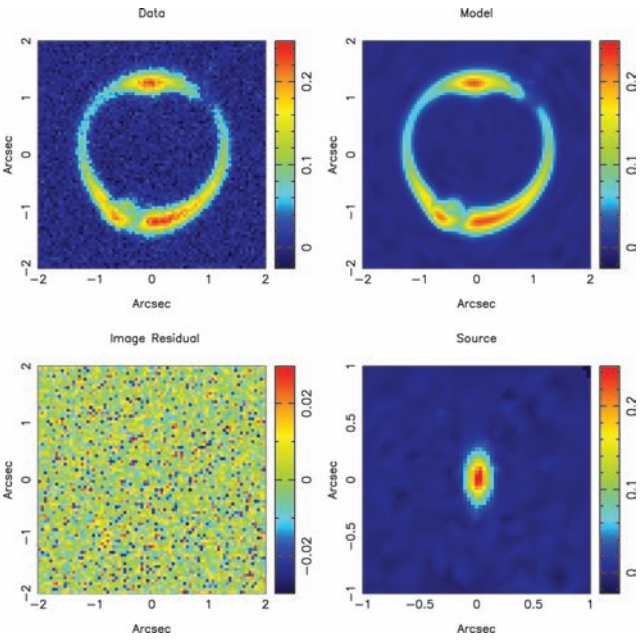

Figure 14. Non-linear reconstruction for the lens L_{13} for a PL + 2NFW model.

a perturbation of $10^7\,\mathrm{M}_\odot$ is readily reconstructed. For these two positions, higher mass models, with the exception of L_2, will not be further analysed. The systems $L_{7,8,9}$ and $L_{10,11,12}$, in which the substructure is located, respectively, inside and outside the ring, represent more difficult scenarios. In the first case, all perturbations

Table 2. Non-linear parameters for the mass model distribution. For each of the considered systems, we report the true set $\boldsymbol{\eta}_{\rm true}$ of non-linear parameters used to create the mock data (true), the best set (best) recovered via the optimization strategy described in Section 3.2 and the average with relative standard deviations values given by the Nested Sampling, under the hypothesis of a single PL potential (PL) and a perturbed PL + NFW lens.

Lens	Model	b (arcsec)	$\sigma_{\rm b}$ (arcsec)	θ (°)	σ_θ (°)	f	σ_f	q	σ_q	log ($\lambda_{\rm s}$)	$\sigma_{\log(\lambda_{\rm s})}$	$m_{\rm sub}$ (10^{10} M$_\odot$)	$\sigma_{m_{\rm sub}}$ (10^{10} M$_\odot$)	$x_{\rm sub}$ (arcsec)	$\sigma_{x_{\rm sub}}$ (arcsec)	$y_{\rm sub}$ (arcsec)	$\sigma_{y_{\rm sub}}$ (arcsec)
L_0	True	1.192		12.23		0.891		0.540									
	Best	1.162		12.35		0.873		0.584		−2.292							
	PL	1.205	0.008	12.15	0.394	0.897	0.005	0.532	0.012	−1.619	0.029						
	PL + NFW	1.187	0.005	14.35	1.907	0.882	0.001	0.538	0.006	−2.221	0.027	0.019	0.013	1.220	1.111	1.103	1.314
L_1	True	1.192		12.23		0.891		0.540				0.001		0.900		1.190	
	Best	1.195		11.87		0.893		0.545		0.912		0.001		0.918		1.174	
	PL	1.197	0.004	10.39	0.253	0.899	0.002	0.543	0.005	0.970	0.039						
	PL + NFW	1.205	0.002	12.85	0.530	0.896	0.001	0.5321	0.003	0.029	0.012	0.004	1.157	0.019	1.436	0.014	
L_2	True	1.192		12.23		0.891		0.540				0.010		0.900		1.190	
	Best	1.213		12.17		0.896		0.522		3.563		0.010		0.917		1.184	
	PL	1.188	0.001	17.81	0.251	0.905	0.001	0.553	0.002	1.187	0.006						
	PL + NFW	1.194	0.004	13.10	0.303	0.892	0.003	0.547	0.005	1.212	0.025	0.013	0.001	0.919	0.008	1.219	0.011
L_4	True	1.192		12.23		0.891		0.540				0.001		−0.500		1.000	
	Best	1.151		11.46		0.874		0.596		3.111		0.001		−0.502		−0.916	
	PL	1.203	0.008	10.87	0.156	0.888	0.004	0.541	0.009	1.107	0.038						
	PL + NFW	1.177	0.006	10.90	0.290	0.877	0.004	0.567	0.007	1.104	0.022	0.0008	0.0003	−0.302	0.096	−0.633	0.019
L_9	True	1.192		12.23		0.891		0.540				0.100		−0.100		−0.600	
	Best	1.186		11.76		0.883		0.559		1.379		0.103		−0.105		−0.595	
	PL	1.251	0.001	21.73	0.018	0.8831	0.0005	0.580	0.001	0.261	0.004						
	PL + NFW	1.215	0.002	11.85	0.284	0.9210	0.0001	0.516	0.004	0.358	0.005	0.9900	0.0002	−0.099	0.001	−0.607	0.001
L_{12}	True	1.192		12.23		0.891		0.540				0.100		−0.900		−1.400	
	Best	1.188		11.73		0.887		0.556		2.831		0.105		−0.919		−1.402	
	PL	1.154	0.029	1.752	0.016	0.881	0.001	0.598	0.027	0.948	0.003						
	PL + NFW	1.203	0.001	11.71	0.297	0.8841	0.0003	0.537	0.002	0.997	0.007	0.101	0.001	−0.906	0.002	−1.409	0.002

below $10^9\,M_\odot$ can be mimicked by an increase in the mass of the main lens within the ring, while in the second case these cannot be easily distinguished from an external shear effect. For the models $L_{1,2,4,9,12}$, convergence is reached after 150 iterations and the perturbations are recovered near their known position (e.g. Figs 8 and 9). The grid-based potential reconstruction indeed leads to a good first estimation for the substructure position.

5.4 Non-linear reconstruction: main lens and substructure

In order to compare with numerical simulations, the mass of the substructure is required. Performing this evaluation with a grid-based reconstruction is more complicated and requires some assumptions (e.g. aperture). To alleviate this problem, we assume a parametric model, in which the substructures are described by a NFW density profile, and we recover the corresponding non-linear parameters, mass and position, using the non-linear Bayesian optimization previously described.

To quantify the mass and position of the substructure and to update the non-linear parameters when a substructure is added, we adopt a multistep non-linear procedure that relatively fast converges to a best PL + NFW mass model. At this level, we neglect the smooth lens L_0, for which a satisfactory model already has been obtained after the first non-linear optimization, and the perturbed models $L_{7,8,10,11}$ for which the substructure could not be recovered. We proceed as follows.

(i) We fix the main lens parameters to the best values found in Section 5.3, $\{\boldsymbol{\eta}_1, \lambda_{s,1}\}$. We set the substructure mass to some guess value. We optimize for the substructure position $\boldsymbol{x}_{\rm sub,1}$.

(ii) We fix $\{\boldsymbol{\eta}_1, \lambda_{s,1}\}$ and the source position $\boldsymbol{x}_{\rm sub,1}$. We optimize for the substructure mass $m_{\rm sub,1}$.

(iii) We run the non-linear procedure described in Section 4 by alternatively optimizing for the main lens, source and substructure parameters and for the level of source regularization.

This leads to a new set of parameters, $\{\boldsymbol{\eta}_{\rm b}, \lambda_{s,b}, m_{\rm sub,b}, \boldsymbol{x}_{\rm sub,b}\}$. Final results for the considered models are listed in Table 2 and the relative residuals are shown in the Figs 4, 6 and 7, respectively. For all the considered lenses, the final reconstruction converges to the noise level.

5.5 Multiple substructures

The lens system L_{13} represents a more complex case in which two substructures are included. In particular, we are interested in testing whether both substructures are detectable and whether their effect may be hidden by the presence of external shear. As for the previously considered cases, we first perform a non-linear reconstruction of the main lens parameters assuming a single PL mass model. For this particular system, we also include the strength $\Gamma_{\rm sh}$ and the position angle $\theta_{\rm sh}$ of the external shear as free parameters. Results for this first step of the reconstruction are shown in Fig. 10. We then run the linear potential reconstruction. One of the two substructures is detected although a significant level of image residuals is left (Fig. 11). The combined effect of external shears ($\Gamma_{\rm sh} = -0.031$) and the substructure in P_1 is not sufficient to explain the perturbation generated by the second substructure at the lowest surface brightness point of the Einstein ring. We therefore include a NFW substructure in the recovered position and run a non-linear reconstruction for the new PL + NFW model, Fig. 12. We are then able to detect also the second substructure, Fig. 13. Finally, we run a global non-linear reconstruction for the PL + 2NFW model (Table 3 and Fig. 14), the noise level is reached and the strength of the external shear is consistent with zero ($\Gamma_{\rm sh} = 0.0001$).

5.6 Nested sampling: the evidence for substructure

When modelling systems as L_0 or $L_{i\geq 1}$, we assume that the best recovered values, under the hypothesis of a single power-law, provide a good description of the true mass distribution and any eventually observed residual could be an indication for the presence of mass substructure. Model comparison within the framework of Bayesian statistics gives us the possibility to test this assumption.

5.6.1 Marginalized Bayesian evidence

In order to statistically compare two models, the marginalized evidence (equation 39) has to be computed. As described in Section 4.1, this multidimensional and non-linear integral can be evaluated using the Nested-Sampling technique by Skilling (2004). Specifically, the two mass models we wish to compare are a single PL, M_0, versus a PL + NWF substructure, M_1. The first one is completely defined by the non-linear parameters ($\boldsymbol{\eta}$, λ_s), while the second needs three extra parameters, namely the substructure mass and position. For all these parameters, prior probabilities have to be defined:

$$P\left(\eta_i\right) = \begin{cases} \text{constant} & \text{for} \quad |\eta_{b,i} - \eta_i| \leq \delta\eta_i \\ 0 & \text{for} \quad |\eta_{b,i} - \eta_i| > \delta\eta_i \end{cases} \tag{48}$$

and

$$P\left(x_{\rm sub,i}\right) = \begin{cases} \text{constant} & \text{for} |x_{\rm sub,b,i} - x_{\rm sub,i}| \leq \delta x_{\rm sub,i} \\ 0 & \text{for} |x_{\rm sub,b,i} - x_{\rm sub,i}| > \delta x_{\rm sub,i} \end{cases} \tag{49}$$

where the elements of $\delta\eta_i$ and $\delta\, x_{\rm sub,i}$ are empirically assessed such that the bulk of the evidence likelihood is included (see Skilling 2004). The prior on the substructure mass is flat between the lower and upper mass limits given by numerical simulations (e.g. Diemand et al. 2007a,b). Given the lenses $L_{0,1,2,4,9,12,13}$, we run the Nested Sampling twice, once for the single PL model and once for the PL + NFW (+NFW) one. The two marginalized evidence with corresponding numerical errors can be compared from Table 4. Despite a certain number of authors suggest the use of Jeffreys' scale (Jeffreys 1961) for model comparison, we adopt here a more conservative criterion. In particular, we note that the perturbed model M_1 for the lens system L_0 is basically consistent with a single smooth PL model M_0, with $\Delta\mathcal{E} \sim 7.85$. The Bayesian factor for the system L_4 is of the order of $\Delta\mathcal{E} \sim 21.5$ in favour of the smooth model M_0, indicating that the detection of such a low-mass substructure can formally not be claimed at a significant level. The reason why we think this substructure is clearly visible in the grid-based results, is that this particular solution is the maximum-posterior (MP) solution, whereas the evidence gives the integral over the entire parameter space. This implies that there must be many solutions near the MP solution that do not show the substructure. This indicates that our approach of quantifying the evidence for substructure is very conservative. On the other hand, the Bayes factor for the lens L_1, $\Delta\mathcal{E} = -17.1$, clearly shows that the detection of a $10^7\,M_\odot$ substructure can be significant when the latter is located at a different position on the ring. Finally, all higher mass perturbations are easily detectable independently of their position relative to the image ring; Bayes factor for L_2, L_9, L_{12} and L_{13} is, in fact, respectively, $\Delta\mathcal{E} = -213.0$, $\Delta\mathcal{E} = -2609.7$, $\Delta\mathcal{E} = -4603.4$ and $\Delta\mathcal{E} = -1835.7$. Substructure properties for these systems are also

Table 3. Non-linear parameters for the mass model distribution for the system L_{13}. We report the true set $\boldsymbol{\eta}_{\rm true}$ of non-linear parameters used to create the mock data (true), the best set (best) recovered via the optimization strategy described in Section 3.2 and the average with relative standard deviations values given by the Nested Sampling, under the hypothesis of a single PL potential (PL) and a perturbed PL + 2NFW lens.

Model	b	$\sigma_{\rm b}$ (arcsec)	θ (arcsec)	σ_θ (°)	f (°)	σ_f	q	σ_q	$\Gamma_{\rm sh}$	$\sigma_{\Gamma_{\rm sh}}$	$\theta_{\rm sh}$	$\sigma_{\theta_{\rm sh}}$ ($10^{10}\,{\rm M}_\odot$)	$\log(\lambda_{\rm s})$ ($10^{10}\,{\rm M}_\odot$)	$\sigma_{\log(\lambda_{\rm s})}$ (arcsec)	$m_{\rm sub}$ (arcsec)	$\sigma_{m_{\rm sub}}$ (arcsec)	$x_{\rm sub}$ (arcsec)	$\sigma_{x_{\rm sub}}$	$y_{\rm sub}$	$\sigma_{y_{\rm sub}}$
True	1.192		12.23		0.891		0.540		0.000		0.000				0.0100		0.900		1.190	
															0.0100		−0.500		−1.000	
Best	1.193		12.32		0.892		0.549		0.0001		0.0001		3.553		0.0100		0.910		1.190	
															0.0100		−0.499		−1.006	
PL	1.182	0.012	12.31	0.022	0.867	0.010	0.580	0.016	−0.001	0.004	0.006	0.020	1.263	0.005						
PL + 2NFW	1.195	0.001	12.32	0.002	0.894	0.015	0.548	0.001	0.0006	0.0002	0.0009	0.0017	1.268	0.003	0.0101	0.0003	0.910	0.002	1.189	0.001
															0.0101	0.0002	−0.499	0.001	−1.000	0.001

Table 4. Marginalized evidence and corresponding standard deviation as obtained via the Nested-Sampling integration. Results are shown for the hypothesis of a smooth lens (PL) and the hypothesis of a clumpy lens potential (PL + NFW).

Lens	Model	$\log \varepsilon$	$\sigma_{\log \varepsilon}$
L_0	PL	26 332.70	0.33
	PL + NFW	26 324.85	0.30
L_1	PL	20 366.86	0.34
	PL + NFW	20 383.95	0.30
L_4	PL	20 292.40	0.33
	PL + NFW	20 270.87	0.29
L_9	PL	17 669.41	0.45
	PL + NFW	20 279.13	0.36
L_{12}	PL	15 786.91	0.33
	PL + NFW	20 390.35	0.37
L_{13}	PL	18 509.76	0.24
	PL + 2NFW	20 346.48	0.49

confidently recovered. The main difference between Jeffreys' scale and our criterion for quantifying the significance level of the substructure detection is observed for the system L_1. If we had to adopt Jeffreys' scale in fact, such detection would have to be claimed decisive while we think it is only significant.

5.7 Posterior probabilities

As discussed in Section 4.1, an interesting byproduct of the Nested-Sampling procedure is an exploration of the posterior probability (equation 34) which provides us with statistical errors on the model parameters (see Table 3). The relative posterior probabilities for $\rm L_0$, L_1 and L_2 are plotted in Figs 15–17, respectively. Lets start by considering the lens system L_0 and the relative probability distribution for the substructure mass. Although the model $\rm M_1$, in terms of marginalized evidence, is consistent with the single smooth PL model $\rm M_0$, there is a small probability for the presence of a substructure with mass up to few $10^8\,{\rm M}_\odot$ located as far as possible from the ring. The effect of such objects on the lensed image would be very small and could be easily hidden by introducing artificial features in the source structure, as suggested by the posterior distributions for the source regularization constant. This means, that from the image point of view, a smooth single PL model and a perturbed PL + NWF with a substructure of $10^8\,{\rm M}_\odot$, located far from ring, are not distinguishable from each other as long as the effect of the perburber can be hidden in the structure of the source. From a probabilistic point of view, however, the second scenario is more unlikely to happen. A similar argument can be applied to the lens L_1 for which a strong degeneracy between the mass and the position of the substructure is observed. We conclude therefore that, although this substructure can be detected at a statistically significant level, its mass and position cannot be confidently assessed yet. In contrast, for systems such as $L_{2,9,12}$, the effect of the substructure is so strong that it cannot be mimicked by the source structure or by a different combination of the substructure parameters. For these cases not only the detection is highly significant, but also the properties of the perturber can be confidently constrained with minimal biases.

6 CONCLUSIONS AND FUTURE WORK

We have introduced a fully Bayesian adaptive method for objectively detecting mass substructure in gravitational-lens galaxies. The implemented method has the following specific features.

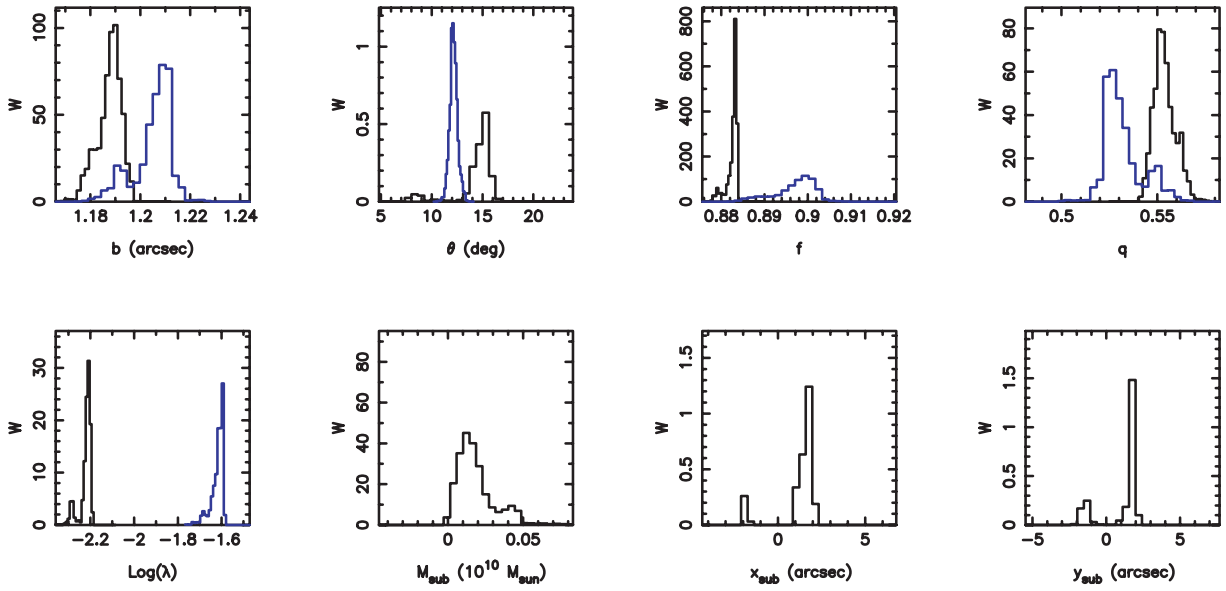

Figure 15. Posterior probability distributions for the non-linear parameters of the smooth lens model L_0 as obtained from the Nested-Sampling evidence exploration. In particular, results for two different models are shown, a smooth PL potential (blue histograms) and a perturbed PL + NFW lens (black histograms). From up left, the lens strength, the position angle, the axis ratio, the slope, the logarithm of the source regularization constant, the substructure mass and position are plotted.

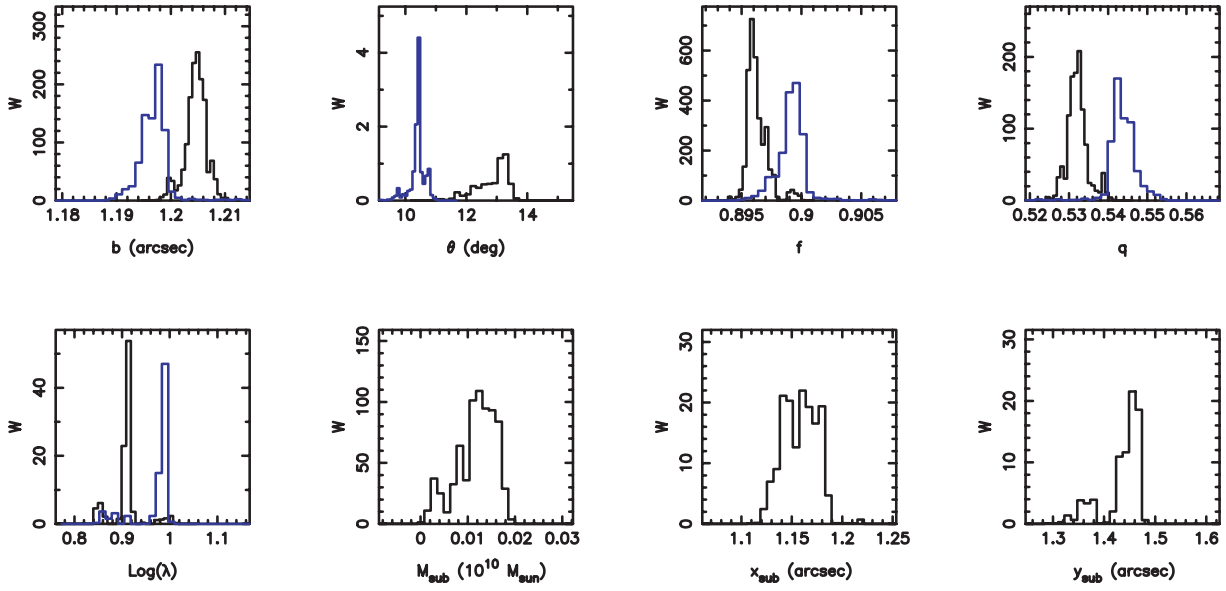

Figure 16. Similar as Fig. 15 for L_1.

(i) Arbitrary imaging data set defined on a regular grid can be modelled, as long as only lensed structure is included. The code is specifically tailored to high-resolution *HST* data sets with a compact PSF that can be sampled by a small number of pixels.

(ii) Different parametric two-dimensional mass models can be used, with a set of free parameter $\boldsymbol{\eta}$. Currently, we have implemented the elliptical PL density models from Barkana (1998), but other models can easily be included. Multiple parametric mass models can be simultaneously optimized.

(iii) A grid-based correction to the parametric potential can iteratively be determined for any perturbation that cannot easily be modelled within the chosen family of potential models (e.g. warps, twists, mass substructures, etc.).

(iv) The source surface brightness structure is determined on a fully adaptive Delaunay tessellation grid, which is updated with every change of the lens potential.

(v) Both model parameter optimization and model ranking are fully embedded in a Bayesian framework. The method takes special care not to change the number of degrees of freedom during the optimization, which is ensured by the adaptive source grid. Methods with a fixed source surface brightness grid cannot do this.

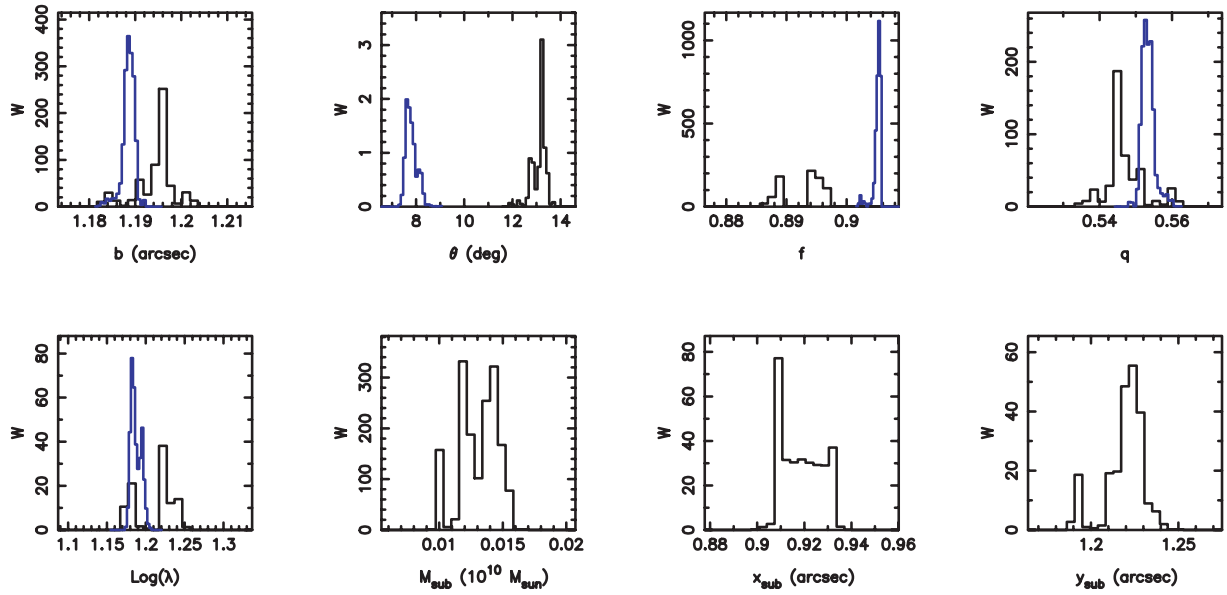

Figure 17. Similar as Fig. 15 for L_2.

(vi) Both source and potential solutions are regularized, based on a smoothness criterion. The choice of regularization can be modified and the level of regularization is set by Bayesian optimization of the evidence. The data itself determine what level of regularization is needed. Hence, overly smooth or overly irregular structure is automatically penalized.

(vii) The MP and the full marginalized probability distribution function of *all* linear and non-linear parameters can be determined, marginalized over all other parameters (including regularization). Hence, a full exploration of *all* uncertainties of the model is undertaken.

(viii) The full marginalized evidence (i.e. the probability of the model given the data) is calculated, which can be used to rank *any* set of model assumptions (e.g. pixel size, PSF) or model families. In our case, we intend to compare smooth models with models that include mass substructure. The marginalized evidence implicitly includes Occam's razor and can be used to assess whether substructure or any other assumption is justified, compared to a null-hypothesis.

The method has been tested and calibrated on a set of artificial but realistic lens systems, based on the lens system SLACS J1627−0055.

The ensemble of mock data consists of a smooth PL lens and 13 clumpy models including one or two NFW substructures. Different values for the mass and the substructure position have been considered. Using the Bayesian optimization strategy developed in this paper, we are able to recover the smooth PL system and all perturbed models with a substructure mass $\gtrsim 10^7\,\mathrm{M}_\odot$ when located at the lowest surface brightness point on the Einstein ring and with a mass $\geq 10^9\,\mathrm{M}_\odot$ when located just inside or outside the ring (i.e. their Einstein rings need to overlap roughly). For all these models, we have convincingly recovered the best set of non-linear parameters describing the lens potential and objectively set the level of regularization.

Furthermore, our implementation of the Nested-Sampling technique provides statistical errors for *all* model parameters and allows us to objectively rank and compare different potential models in terms of Bayesian evidence, removing as much as possible any subjective choices. Any choice can quantitatively be ranked. For each of the lens systems, we compare a complete smooth PL mass model with a perturbed PL + NFW (+NFW) one. The method here developed allows us to solve simultaneously for the lens potential and the lensed source. The latter, in particular, is reconstructed on an adaptive grid which is recomputed at every step of the optimization, allowing to take into account the correct number of degrees of freedom.

In this paper, we have considered systems which contain at most two CDM substructures. Although it may appear as a very small number when compared with predictions from *N*-body simulations within the virial radius, this represents a realistic scenario. As we have shown, our method, with current *HST* data, is mostly sensitive to perturbations with mass $\gtrsim 10^7\,\mathrm{M}_\odot$ and located on the Einstein ring ($\Delta\theta \sim \mu\theta_{\mathrm{ER}}$). The projected volume that we are able to probe is therefore small compared to the projected volume within the virial radius. The probability that more than two substructures have this right combination of mass and position is relatively low and we expect most of the real systems to be dominated by one or at most two perturbers. Despite these new results, further improvements can still be made. We think, for example, that an adaptive source grid based on surface brightness, rather than magnification, or a combination, could be more suitable for the scientific problem considered here.

The method will soon be applied to real systems, as for example from the *Sloan Lens ACS* sample of massive early-type galaxies (Bolton et al. 2006; Koopmans et al. 2006; Treu et al. 2006). This will lead to powerful new constraints or limits on the fraction and mass distribution of substructure. Results will be compared with CDM simulations.

ACKNOWLEDGMENTS

The authors would like to thank Matteo Barnabè, Oliver Czoske, Antonaldo Diaferio, Phil Marshall, Sherry Suyu and the anonymous referee for useful discussions. They also thank the Kavli Institute for Theoretical Physics for hosting the gravitational lensing workshop in fall 2006, during which important parts of this work were

made. SV and LVEK are supported (in part) through an NWO-VIDI program subsidy (project number 639.042.505).

REFERENCES

Barkana R., 1998, ApJ, 502, 531
Barnabè M., Koopmans L. V. E., 2007, ApJ, 666, 726
Barnes J., 1992, ApJ, 393, 484
Bergström L., Edsjö J., Gondolo P., Ullio P., 1999, Phys. Rev. D, 59, 043506
Binney J., Evans N. W., 2001, MNRAS, 327, L27
Blandford R., Narayan R., 1986, ApJ, 310, 568
Bolton A. S., Burles S., Koopmans L. V. E., Treu T., Moustakas L. A., 2006, ApJ, 638, 703
Bradač M., Schneider P., Steinmetz M., Lombardi M., King L. J., Porcas R., 2002, A&A, 388, 373
Bradač M., Schneider P., Lombardi M., Steinmetz M., Koopmans L. V. E., Navarro J. F., 2004, A&A, 423, 797
Brewer B. J., Lewis G. F., 2006, ApJ, 637, 608
Bullock J. S., Kravtsov A. V., Weinberg D. H., 2000, ApJ, 539, 517
Burkert A., 1995, ApJ, 447, L25
Burles S., Nollett K. M., Turner M. S., 2001, Phys. Rev. D, 63, 063512
Calcáneo-Roldán C., Moore B., 2000, Phys. Rev. D, 62, 123005
Cen R., 2001, ApJ, 546, L77
Chen J., Rozo E., Dalal N., Taylor J. E., 2007, ApJ, 659, 52
Chiba M., 2002, ApJ, 565, 17
Colafrancesco S., Profumo S., Ullio P., 2006, A&A, 455, 21
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Colin P., Avila-Reese V., Valenzuela O., 2000, ApJ, 542, 622
Croft R. A. C., Weinberg D. H., Bolte M., Burles S., Hernquist L., Katz N., Kirkman D., Tytler D., 2002, ApJ, 581, 20
Dalal N., Kochanek C. S., 2002, ApJ, 572, 25
de Bernardis P. et al., 2002, ApJ, 564, 559
de Blok W. J. G., Bosma A., 2002, A&A, 358, 816
de Blok W. J. G., McGaugh S. S., Rubin V. C., 2001, AJ, 122, 2396
Diemand J., Kuhlen M., Madau P., 2007a, ApJ, 667, 859
Diemand J., Kuhlen M., Madau P., 2007b, ApJ, 657, 262
Dye S., Warren S. J., 2005, ApJ, 623, 31
Efstathiou G. P. et al., 2002, MNRAS, 330, L29
Frenk C. S., White S. D. M., Davis M., Efstathiou G. P., 1988, ApJ, 327, 507
Goodman J., 2000, New Astron., 5, 103
Hamilton A. J. S., Tegmark M., 2002, MNRAS, 330, 506
Ibata R. A., Lewis G. F., Irwin M. J., Quinn T., 2002, MNRAS, 332, 915
Jaffe A. H. et al., 2001, Phys. Rev. Lett., 86, 3475
Jeffreys H., 1961, Theory of Probability, 3rd edn. Oxford Univ. Press, Oxford
Kamionkowski M., Liddle A., 2000, Phys. Rev. Lett., 84, 4525
Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201
Keeton C. R., Gaudi B. S., Petters A. O., 2003, ApJ, 598, 138
Keeton C. R., Gaudi B. S., Petters A. O., 2005, ApJ, 635, 35
Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, ApJ, 522, 82
Kochanek C. S., Dalal N., 2004, ApJ, 610, 69
Komatsu E. et al., 2008, ApJS, preprint (arXiv:0803.0547)
Koopmans L. V. E., 2005, MNRAS, 363, 1136
Koopmans L. V. E., Treu T., 2002, ApJ, 568, L5
Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006, ApJ, 649, 599
Kravtsov A. V., Gnedin O. Y., Klypin A. A., 2004, ApJ, 609, 482
Kuzio de Naray R., McGaugh S. S., de Blok W. J. G., Bosma A., 2006, ApJS, 165, 461
Macciò A. V., Miranda M., 2006, MNRAS, 368, 599
McGaugh S. S., de Blok W. J. G., 1998, ApJ, 499, 41
McGaugh S. S., Barker M. K., de Blok W. J. G., 2003, ApJ, 584, 566
MacKay D. J. C., 1992, PhD thesis, Caltech
MacKay D. J. C., 2003, Information Theory, Inference and Learning Algorithms, Cambridge Univ. Press, Cambridge
McKean J. P. et al., 2007, MNRAS, 378, 109
Mao S., Schneider P., 1998, MNRAS, 295, 587
Mao S., Jing Y., Ostriker J. P., Weller J., 2004, ApJ, 604, L5
Mayer L., Moore B., Quinn T., Governato F., Stadel J., 2002, MNRAS, 336, 119
Metcalf R. B., Zhao H., 2002, ApJ, 567, L5
Moore B., 1994, Nat, 370, 629
Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, ApJ, 524, L19
Moore B., Calcáneo-Roldán C., Stadel J., Quinn T., Lake G., Ghigna S., Governato F., 2001, Phys. Rev. D, 64, 063508
Moore B., Diemand J., Madau P., Zemp M., Stadel J., 2006, MNRAS, 368, 563
More A., McKean J. P., More S., Porcas R. W., Koopmans L. V. E., Garret M. A., 2008, MNRAS, submitted (arXiv:0810.5341)
Percival W. J. et al., 2001, MNRAS, 327, 1297
Phillips J., Weinberg D. H., Croft R. A. C., Hernquist L., Katz N., Pettini M., 2001, ApJ, 560, 15
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in FORTRAN. The Art of Scientific Computing, 2nd edn. Cambridge Univ. Press, Cambridge
Reiss D. J., Germany L. M., Schmidt B. P., Stubbs C. W., 1998, AJ, 115, 26
Rhee G., Valenzuela O., Klypin A., Holtzman J., Moorthy B., 2004, ApJ, 617, 1059
Riotto A., Tkachev I., 2000, Phys. Lett. B, 484, 177
Ros E., Guirado J. C., Marcaide J. M., Perez-Torres M. A., Falco E. E., Munoz J. A., Alberdi A., Lara A., 2000, A&A, 362, 845
Simon J. D., Bolatto A. D., Leroy A., Blitz L., 2003, ApJ, 596, 957
Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, AIP Conf. Series Vol. 735. Am. Inst. Phys., New York
Spergel D. N., Steinhardt P. J., 2000, Phys. Rev. Lett., 84, 3760
Spergel D. N. et al., 2003, ApJS, 148, 175
Stoehr F., White S. D. M., Springel V., Tormen G., Yoshida N., 2003, MNRAS, 345, 1313
Suyu S. H., Blandford R. D., 2006, MNRAS, 366, 39
Suyu S. H., Marshall P. J., Hobson M. P., Blandford R. D., 2006, MNRAS, 371, 983
Tonry J. L. et al., 2003, ApJ, 594, 1
Toomre A., 1977, in Tinsley B. M., Larson R. B., eds, Evolution of Galaxies and Stellar Populations Mergers and Some Consequences. Yale Univ. Obs., New Haven, p. 401
Treu T., Koopmans L. V. E., 2004, ApJ, 611, 739
Treu T., Koopmans L. V. E., Bolton A. S., Burles S., Moustakas L. A., 2006, ApJ, 640, 662
Trotter C. S., Winn J. N., Hewitt J. N., 2000, ApJ, 535, 671
Warren S. J., Dye S., 2003, ApJ, 590, 673
White S. D. M., Frenk C. S., 1991, ApJ, 379, 52
Zakharov A. F., 1995, A&A, 293, 1
Zentner A. R., Bullock J. S., 2003, ApJ, 598, 49

This paper has been typeset from a TEX/LATEX file prepared by the author.